Join Cost for Unit Selection Speech Synthesis

Jithendra Vepa

Thesis submitted for the degree of Doctor of Philosophy

University of Edinburgh

Jan 2004

Dedicated

To

My Very Dear

Parents and Grandparents

Acknowledgements

I owe many thanks to my supervisor Simon King for his invaluable guidance and timely advices throughout my PhD. Without his support I could not have finished my thesis. I also thank him for spending long hours reading my thesis and correcting my non-native English.

I thank Paul Taylor, my second supervisor, for his encouragement and support. Many thanks to Rhetorical Systems Ltd. and Centre for Speech Technology Research (CSTR) for their financial support. I also thank Rhetorical for providing their speech databases and rVoice speech synthesis system.

I thank one and all at CSTR. Firstly, many thanks to Joe Frankel for allowing me to use his LDM code, patiently answering my questions and proof reading one chapter of my thesis. I also thank Korin Richmond for his suggestions and helpful comments while proofing the text. Many thanks to Rob Clark for his help in fixing many software problems. I thank Yoshinori Shiga for interesting discussions from speech signal processing to Indian food. Many thanks to James Horlock and Mirjam Wester for their good company.

I thank all participants in my listening tests for their valuable time and patience. I also thank Peter Rutten, Mathew Aylett, Justin Fackrell and Alexander Gutkin of Rhetorical for helping me out with rVoice system.

I am forever indebted to my parents for their efforts and devotion to give their children the very best of all. Without their love and inspiration I would never have got this far. I also thank my sisters Janaki and Madhavi and my brother Srinivas for their unconditional support and encouragement. Finally, many thanks to my wife, Jahnavi for sharing with me the moments of excitement as well as frustration during last few months.

Declaration

I declare that, apart from where properly indicated, the work contained in this thesis is entirely the product of my own work.

(Jithendra Vepa)

Abstract

Undoubtedly, state-of-the-art **unit selection-**based concatenative speech synthesis systems can produce very high quality synthetic speech. This is due to a large speech database containing many instances of each speech unit, with a varied and natural distribution of prosodic and spectral characteristics.

The **join cost**, which measures how well two units can be joined together is one of the main criteria for selecting appropriate units from this large speech database. The ideal join cost is one that measures perceived discontinuity based on easily measurable spectral properties of the units being joined, in order to ensure smooth and natural-sounding synthetic speech.

During the first part of my research, I have investigated various spectrally-based distance measures for use in computation of the join cost by designing a perceptual listening experiment. A variation on the usual perceptual test paradigm is proposed in this thesis by deliberately including a **wide range of qualities** of join in **polysyllabic words**. The test stimuli are obtained using a state-of-the-art unit selection text-to-speech system: *rVoice* from Rhetorical Systems Ltd. Three spectral features – Mel-frequency cepstral coefficients (**MFCC**), line spectral frequencies (**LSF**) and multiple centroid analysis (**MCA**) parameters – and various statistical distances – Euclidean, Kullback-Leibler, Mahalanobis – are used for obtaining distance measures. Based on the correlations between perceptual scores and these spectral distances, I propose new spectral distance measures which have good correlation with human perception of concatenation discontinuities.

The second part of my research concentrates on combining join cost computation and the smoothing operation, which is required to disguise joins, by learning an underlying representation from the acoustic signal. In order to accomplish this task, I have chosen linear dynamic models (LDM), sometimes known as **Kalman filters**. Three different initialisation schemes are used prior to Expectation-Maximisation (EM) in LDM training. Once the models are trained, the join cost is computed based on the error between model predictions and actual observations. Analytical measures are derived based on the shape of this error plot. These measures and initialisation schemes are compared by computing correlations using the perceptual data. The LDMs are also able to smooth the observations which are then used to synthesise speech.

I have compared the best three join cost functions, chosen from the first and second parts of my research, subjectively using a listening test in the third part of my research. In this test, I also evaluated different smoothing methods: no smoothing, linear smoothing and smoothing achieved using LDMs.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Speech synthesis

The automatic generation of speech from any text, whether it was directly introduced in the computer by human or scanned and submitted to an Optical Character Recognition (OCR) system is referred to as Text-to-Speech (TTS) synthesis (Dutoit 1997). A typical TTS system has two sets of modules (Klatt 1987), as shown in Figure 1.1.

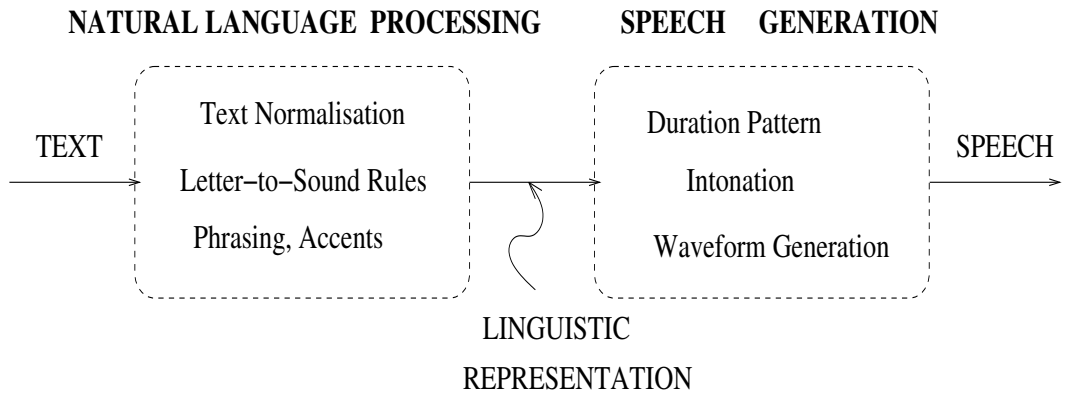

Figure 1.1: A typical TTS system

The first set of modules analyses the text to determine the underlying structure of the sentence and phonetic composition of each word. This involves tokenisation, i.e. breaking the input text into sentences and then to words. Then abbreviations, punctuations and acronyms are identified and expanded into full text. Finally, using a lexicon and letter-to-sound rules, the words are converted into their corresponding

phonetic transcription.

The second set of modules transforms this abstract linguistic representation into an output speech signal. Initially, they generate prosody by predicting pitch contours and durations for the utterance. Then, they produce the synthetic waveform using one of a number of techniques. The various methods of speech synthesis will be discussed in the following section.

1.1.1 Methods for waveform generation

The methods available for speech synthesis fall into two categories, based on whether a form of speech production model is used: model-based (rule-based) and concatenative systems.

1.1.1.1 Model-based systems

These systems have some sort of simplified model of speech production, typically driven by rules. Hence, they are also termed **rule-based** systems. These rules may be purely heuristics in the acoustic domain or based on a model of the larynx and vocal tract. There are two main steps in this method: generating a set of rules and synthesising the speech waveform. First, the rules map from phonetic inputs to acoustic parameters, such as formants, anti-formants and bandwidths. Then, these parameters are sent to a synthesiser, typically a formant synthesiser, to generate a digital speech signal.
Example Systems: **MITalk** (Allen, Hunnicutt, Carlson & Gramström 1979, Allen, Hunnicutt & Klatt 1987), **Klattalk** (Klatt 1987), **DECTalk** (Klatt 1987)

Advantages:

- *Allow incorporation of linguistic knowledge*
- *Low memory requirements*
- *Ability to tweak the rules to gradually refine them*

Disadvantages:

- *Quality is inherently limited by the model of speech production*

- *Formant frequencies and bandwidths are difficult to estimate*
- *Hard to transfer to a new language, since large parts need to be entirely re-written*

1.1.1.2 Concatenative systems

In contrast to the model-based approach, concatenative systems use pre-recorded segments of real speech from a single speaker. During synthesis, the chosen segments are concatenated to generate synthetic speech.
Example Systems: **Festival** (Black & Taylor 1997*b*, Taylor, Black & Caley 1998), **CHATR** (Campbell 1996), **Next-Gen** (Beutnagel, Conkie, Schroeter, Stylianou & Syrdal 1999), **RealSpeak** (Coorman, Fackrell, Rutten & van Coile 2000)
Advantages:

- *Can change the voice without changing software*
- *Can produce very natural sounding synthetic speech*

Disadvantages:

- *Large database of speech required*
- *Need large memory and processing power*
- *Can sometimes hear the joins*

Concatenative speech synthesis systems need an inventory of speech segments (units) to concatenate and produce synthetic speech. The choice of unit for concatenation is an open issue. Though using larger units can produce higher quality with fewer joins, this will exponentially increase the size of the inventory. Shorter units result in a smaller inventory, but more joins and therefore more concatenation artifacts in synthesis. Hence, there is a trade-off between unit size and number of joins.

The smallest possible units for TTS would be phones, and the concatenation points would be phone boundaries. However, these are acoustically unstable points due to coarticulatory effects between adjacent phones. Hence, in order to make the joins at mid-phone positions, the smallest unit used in current systems is now the **diphone**,

which extends from the middle of one phone to the middle of the next phone. The first successful concatenative systems were diphone-based (Black & Taylor 1997*b*).

1.1.2 Diphone synthesis

As the name implies, this method uses diphones as the speech database units for concatenation. The diphones involve the transitions between phones and have acoustically stable boundaries. Hence, they are less affected by contextual effects like co-articulation and assimilation.

To implement this method, first we need to build the inventory, which consists of one copy of every possible diphone in the language. The maximum inventory size is the square of the number of phonemes. Since diphones are reasonable in number (around 1600 for English) they have formed the basis for most concatenation systems (Klatt 1987, page 758). There are various free softwares packages available, such as FestVox[1] to assist in making this diphone database.

Once the database is ready, then synthesis is mere concatenation of the required diphones. However, this does not produce natural speech as these diphones lack the correct duration and pitch. Most diphone systems, such as Festival (Black & Taylor 1997*b*) use nonsense carrier words to collect diphones (Lenzo & Black 2000). Thus, there is a need for modifying the prosody, whilst generating synthetic output. In recent years many techniques, such as TD-PSOLA (Charpentier & Stella 1986, Hamon, Moulines & Charpentier 1989), LPC (Makhoul 1973, Makhoul 1975), RELP (Residual excited LPC) (Hunt, Zwierynski & Carr 1989) and MBROLA (Dutoit & Leich 1993), have been proposed for generating waveforms, all of which also facilitate prosody modifications.

1.1.2.1 Pros and cons of diphone synthesis

- Produces intelligible speech
- Easily adaptable for other voices and other languages
- But not natural, often hear the joins due to discontinuities at concatenation points

[1] www.festvox.org

  - *Can use advanced signal processing techniques for better prosody and smooth concatenation, such as residual LP-PSOLA, MBROLA, but they involve more computation and may add artifacts*

- Does not handle other prosodic factors, such as distinctions between stressed and unstressed syllables properly

  - *Though this could be overcome by designing a diphone inventory with different versions of each diphone, it would be cumbersome*

1.1.3 Unit selection-based concatenative speech synthesis

Unit selection-based speech synthesis systems have become popular recently because of their highly natural-sounding synthetic speech (Hunt & Black 1996, Beutnagel et al. 1999, Donovan & Eide 1998, Huang, Acero, Adcock, Hon, Goldsmith, Liu & Plumpe 1996, Coorman et al. 2000). These systems have large speech databases containing many instances of each speech unit (e.g. diphone), with a varied and natural distribution of prosodic and spectral characteristics. When synthesising an utterance, the selection of the best unit sequence from the database is based on a combination of two costs: target cost (how closely candidate units in the inventory match the required targets) and join cost (how well neighbouring units can be joined) (Hunt & Black 1996).

1.1.3.1 Target cost

An estimate of the difference between a candidate unit (in the database) and the target unit is known as target cost. This is computed as the weighted sum of the differences between various features of target and candidate units, such as:

- *Duration*
- *F0*
- *Stress*
- *Word (orthographic match)*
- *Syllable position*

- *Utterance position*
- *Position in a word*

1.1.3.2 Join (concatenation) cost

This is a measure of the perceptual quality of a join between consecutive units. This may also determined as the weighted sum of sub-costs, such as:

- *Differences in F0 and amplitude*
- *Mismatch in different spectral parameters, like*
  - **MFCC** (Mel Frequency Cepstral Coefficients) (Rabiner & Juang 1993, Huang, Acero & Hon 2001)
  - **LPC** (Linear Predictive Coding Coefficients) (Makhoul 1973, Makhoul 1975)
  - **LSFs** (Line Spectral Frequencies) (Itakura 1975, Soong & Juang 1984)
  - **MCA** (Multiple Centroid Analysis) (Crowe & Jack 1987)

A detailed description of above speech parameterisations will be presented in chapter 2 (section 2.3).

1.1.3.3 Viterbi search

The selection of the best unit sequence is a compromise between choosing units with low target and low join costs. In other words we are looking for the optimal path sequence in the candidate lattice. For this purpose, a popular technique based on dynamic programming, known as the Viterbi algorithm (Viterbi 1967) can be used.

In order to employ this Viterbi search, we need to construct a lattice of the candidate lists by making a path between each possible pair of nodes (boundaries of candidate units). Figure 1.2 shows the lattice of candidate units with the target and join costs. The Viterbi search moves left to right through this lattice, and in doing so it calculates a partial path (cumulative) cost. This is the sum of the target and join costs of units in a given path. As the search goes further, the Viterbi algorithm picks and remembers the best path, so far. Hence, for a given candidate unit, it uses only the lowest cost path

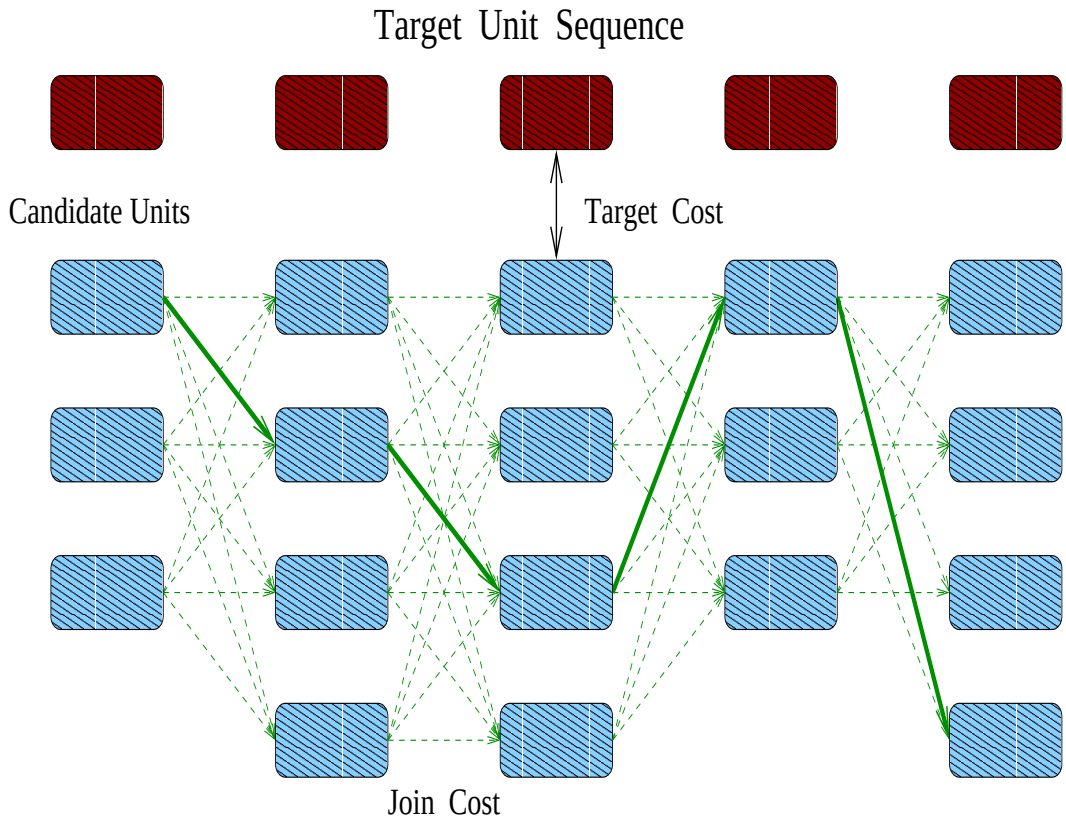

Figure 1.2: Lattice of candidate units in Viterbi search; all possible paths are shown in dashed lines and the solid line shows the optimal path.

till that point. Once, the search terminates, the units forming the path with the lowest overall cost are selected.

1.1.3.4 Unit selection vs Diphone synthesis

- Undoubtedly, unit selection can produce synthetic speech with very high naturalness, compared to diphone synthesis (Iwahashi, Kaiki & Sagisaka 1992, Campbell 1996, Hunt & Black 1996, Huang et al. 1996, Beutnagel et al. 1999)
- Unit selection does not require much prosody modification due to its large database of speech units with varied prosodic features
- Less signal processing is required in unit selection compared to diphone synthesis
- Unit selection synthesis performs better in many domains, like weather forecasting, sporting news or dialogue systems since it can be made more domain specific (Black 2002)
- Diphone synthesis requires a much smaller database and less computing power

1.2 Literature review

A comprehensive review of research publications in the area of speech synthesis with an emphasis on unit selection-based concatenative speech synthesis now follows. We also review linear dynamic models, which we used for computing join cost and smoothing (chapters 5 and 6). This literature survey is categorised into the following topics:

- Unit selection-based speech synthesis systems
- Join cost functions and objective distance measures for spectral discontinuities
- Linear dynamic models for speech modelling

Before starting with each of the above categories, it is worth mentioning the review paper on *text-to-speech conversion for English* by Dennis H. Klatt (1987). This review traces the early work from the thirties on the development of speech synthesisers, the

evolution of phonemic rules and formulation of techniques for text analysis. Various strategies for doing synthesis, such as formant-based rule synthesis, articulatory synthesis and waveform concatenation are also described. In this review, the main emphasis is on synthesis-by-rule programs that drive formant synthesisers and many such example systems are presented. Numerous issues are highlighted to improve text-to-speech systems, in order to achieve perfect human-sounding speech. Finally, an extensive review of various synthesisers from both industrial and academic research is presented.

1.2.1 Unit selection

In the late 1980s, researchers at ATR, Japan developed the first concatenative speech synthesis system using a large speech database instead of using a simple diphone inventory. This section starts by citing some of their work in this context. Then, the description of various unit selection synthesis systems developed at well known research laboratories will follow.

Iwahashi et al. (1992) proposed a new scheme for the selection of segments for concatenative synthesis based on minimising spectral distortions between a target utterance and candidate segments. The spectral distortion measures used in this study are:

- *Contextual spectral difference (CSD) of source and target contexts (computed as the Euclidean distance between the centroids of the sets of time-normalised cepstral parameters of the segment in each of two contexts)*
- *Prototypicality of a segment, also calculated as the Euclidean distance between cepstral parameters of segment and centroid of the set of segments of same context*
- *Disconcatability of phoneme, measure of quality degradation due to concatenation*
- *Acoustic continuity between the concatenated segments, calculated as the spectral distance between two segment boundaries*

They also described a search method for selecting segments from a large speech database. Their perceptual test showed that CSD and acoustic continuity at the segment boundary are good measures for improving synthetic speech quality.

1.2.1.1 CHATR

The work done at ATR, Japan with respect to concatenative synthesis, led to the development of the CHATR speech synthesis system (Black & Taylor 1994, Campbell 1996). This system produces synthetic speech without using signal processing, instead using re-sequencing of carefully selected phone-sized segments from a pre-recorded speech corpus.

First, it involves creation of an index of phones and their prosodic characteristics for each utterance in the corpus. CHATR uses both prosodic and phonetic features of units, contrary to its predecessor, the ν-talk system, which was based only on minimising acoustic distortions between selected units and the target spectrum (Iwahashi et al. 1992). Then, it determines the optimal sequence of segments from a speech corpus which gives the best approximation to a desired utterance. To find the optimal sequence of segments for concatenation, the synthesiser selects amongst candidates in the database using a weighted combination of their acoustic and prosodic features to maximise continuity between segments while at the same time minimising the distance of each from its prosodic target. To our knowledge, this is the first unit selection synthesiser using a large speech corpus.

Unit selection framework An excellent paper by Hunt & Black (1996) describes a widely used framework for unit selection-based concatenative speech synthesis with two cost functions. They proposed that speech units in a synthesis database can be considered as a state transition network in which the state occupancy cost (**target cost**) is the distance between a database unit and a target. The transition cost (**concatenation cost**) is an estimate of the quality of the concatenation of two consecutive units. Then, optimal unit selection is performed using a Viterbi search after pruning the database first with phonetic contexts of a target and then based on target and concatenation costs. They pointed out similarities of this framework to HMM-based speech recognition systems, the important distinction being that Markov models are probabilistic, whereas this framework uses cost functions.

They also presented two methods – weight space search and regression training – for training weights of the target and concatenation (join) cost functions. In weight space

search, the best weight set is chosen among a range of weight sets based on the objective distances between natural waveforms and synthetic waveforms obtained from a sequence of units in the database. The regression determines the weights for the concatenation cost and the weights for the target cost separately using a training algorithm, which treats the example unit as a target unit and computes the acoustic differences of a target and all other instances of the same unit in the database. Finally, they observed that both these weight training methods provide better quality synthesis than hand-tuned weights. Specifically, the regression training is more effective than weight space search due to its flexibility as well as its fast computation. This research was carried out within the CHATR speech synthesis system.

Further work on use of new acoustic features in the unit selection of CHATR was done by Ding & Campbell (1997). They used voice source parameters, formants and bandwidths for join cost computation instead of cepstral coefficients, previously used in the CHATR system (Hunt & Black 1996). These features were estimated automatically from speech waveforms using the ARX joint estimation method (Ding, Kasuya & Adachi 1995). They performed both objective and subjective experiments to verify the validity of these new features. The objective distance measure was based on these glottal source and formant parameters. Earlier, the CHATR system used cepstral distance between the waveforms as the objective distance measure (Black & Campbell 1995). Their results showed that these new acoustic features outperformed the previous ones, i.e. cepstral parameters, and also produced natural speech.

1.2.1.2 Festival

The Festival speech synthesis system (Black & Taylor 1997*b*) was first developed at CSTR, Edinburgh for synthesising speech from a diphone inventory. This system provides a flexible and modular architecture for easy experimentation and competitive evaluation of different modules (Taylor et al. 1998). Hence, it became a benchmark for many speech synthesis systems.

Black & Taylor proposed a new method of unit selection-based on clustering similar units for unit selection synthesis (Black & Taylor 1997*a*). In this approach, a large unit inventory was created by automatically clustering units of a particular phone depending

on their phonetic and prosodic context. Thus for each phone, they built a decision tree using the classification and regression trees (CART) (Brieman, Friedman, Olshen & Stone 1992) method. The target cost was simply computed as the distance of a unit to its cluster centre.

An optimal coupling technique (Conkie & Isard 1996) was used to measure the concatenation cost and also to select a position for the join. Since they used phone size units, this technique helped them to select more stable positions towards the centre of the phone for the join. Finally, Viterbi search was used to find the optimal path through the candidate units. They carried out listening tests to judge the quality of synthetic speech and observed that a cluster size of around 10 produced the best quality.

Phonological structure matching (PSM) Another new technique for unit selection speech synthesis was presented in (Taylor & Black 1999, Taylor 2000). In this method, both sentences in the speech database and the synthesis target are represented as phonological trees. The phonological tree is constructed by combining the metrical tree for the sentence with the sub-syllabic phonological structure. Metrical trees are binary branching trees whose nodes have relative metrical strength relations.

In this technique, first we need to generate a **target tree** representing the phonological features of the utterance to be synthesised. Then, unit selection and synthesis can be done in three steps. The first step is finding the candidates by comparing nodes in the target tree to the nodes in the database. In the second step, the best unit sequence has to be chosen from a target tree containing multiple candidates. In order to do this, PSM uses two costs, target and concatenation costs, as in the unit selection framework described by Hunt & Black (1996). Once the best units are selected, synthetic speech is produced by concatenating them.

The PSM method offers many advantages over traditional unit selection approach. The most important one being that it tries to select longer units and thus reduces the number of joins. Hence, it is possible to achieve high quality synthesis. In this method only a few parameters or weights needed to be set by training. Also, the computation requirements for this method are less compared to traditional unit selection approach presented in (Hunt & Black 1996). However, the success of this algorithm

mainly depends on accurate representation of phonological structure. Still, it did not overcome common problems that occur in using cost functions and their construction. Considering its strengths and weaknesses, it is well suitable for domain specific tasks as building this type of systems is easy and this method also finds longer units.

1.2.1.3 Laureate

British Telecom (BT) developed a text-to-speech system known as Laureate (Breen & Jackson 1998*b*, Breen & Jackson 1998*a*), which selects units purely based on phonological features. The speech database used in Laureate contains at least one instance of every diphone and is composed of phonetically rich passages. The unit selection process is not restricted to diphones and rather it selects N-phone (usually, triphone) units.

Once the recorded speech is annotated, *phoneme context trees* are built. The context tree has three levels, the first level contains indexes to particular phonemes. The second level records triphones and the third level stores one extra phoneme on either side of the triphone. The length of context tree window specifies the maximum unit size.

The first stage of unit selection examines the database for the candidate units: left diphone, right diphone, triphone and centre phone. For each context window, all or a subset of these candidate units are selected and placed into a work space. Then the distance metric is used to select the units which most closely match the target context using nominal features of the phonemes within the unit and its neighbours and also features associated with linguistic environment of the unit. In the second stage, it uses dynamic programming to select the best possible path through the work space based on a cost, which is lower for overlapping units than adjoining units. Finally, the output of the unit selection process is used for generating synthetic speech.

1.2.1.4 Next-Gen

AT&T Labs developed a TTS system for general US English text based on Festival, CHATR and its own Flextalk systems, known as the Next-Gen TTS system (Beutnagel et al. 1999). This system was implemented within the Festival framework due to its high degree of flexibility. Text processing components were borrowed from AT&T Bell

Labs' previous system, Flextalk. Then, unit selection was done based on the CHATR unit selection method, except using half-phones as units instead of phonemes. Finally, for synthesising speech it used a harmonic plus noise model (HNM) (Stylianou, Dutoit & Schroeter 1997, Stylianou 2001).

Formal listening tests were conducted to evaluate the system. These tests indicated that a suitable speaker can have a very significant effect on voice ratings. In other words, selecting an appropriate voice for constructing the speech inventory is the most important task. The best performance was achieved without any prosody modifications during synthesis except some smoothing at unit boundaries. The predicted prosody was still used as a key to find the optimal sequence of units from the inventory, but was not imposed on the selected units. Also, these tests revealed that pruning the database without sacrificing the quality was not easy.

Conkie, Beutnagel, Syrdal & Brown (2000) proposed two methods to reduce the complexity of the unit selection in the above system, and thus accomplish fast unit selection without any degradation in the synthesis quality. The first pre-selection method involves pre-computing not the exact set of units that are relevant for each specific context, but a relatively small set of all the units that may be relevant for a closely related group of contexts. The second method uses a **join cost cache** and this led to more significant speed-ups in synthesis than the first method. Their formal listening tests confirmed that pre-selection did not degrade the quality of synthetic speech.

1.2.1.5 Whistler

A trainable text-to-speech system, called Whistler (Whisper Highly Intelligent Stochastic TaLkER) was developed at Microsoft (Huang et al. 1996). Basic text analysis was carried out using the Microsoft NLP engine. Whistler used data-driven probabilistic modelling, which yielded a relatively natural sounding prosody in synthetic speech. Prosody modification was implemented using pitch-synchronous LPC parameters and their residual waveforms.

To achieve more natural synthetic speech, Whistler uses decision tree based senones as the synthesis units. A senone is a context-dependent sub-phonetic unit which is equivalent to a HMM state in a triphone. These senone decision trees were generated

from the database. Then, the selection of good unit instances was done by using an objective function based on HMM scores. During run-time, the synthesiser dynamically selects the best senone instance sequence which minimises the spectral distortion at the join.

In addition, Whistler tried to extract senone units which form longer units, and thus improve the synthesis quality by minimising distortions due to concatenation. The experiments performed on this system, confirmed that multiple instance based senone synthesiser produces more natural sounding and better quality synthetic speech than a single instance diphone synthesiser.

1.2.1.6 IBM trainable

Another trainable speech synthesis system was developed by Donovan et al at IBM (Donovan, Ittycheriah, Franz, Ramabhadran, Eide, Viswanathan, Bakis, Hamza, Picheny, Gleason, Rutherfoord, Cox, Green, Janke, Revelin, Waast, Zeller, Guenther & Kunzmann 2001, Donovan & Eide 1998). This system uses a set of speaker-dependent decision-tree state-clustered HMMs to automatically generate a leaf level segmentation of a large single-speaker database. During synthesis, text is first converted into a phone sequence, then the decision trees are used to convert this phone sequence into an acoustic, duration, and energy leaf sequence. Duration, energy and pitch are predicted using separate trainable models.

To determine the optimal segment sequence from the available acoustic decision tree leaves for synthesis, a dynamic programming search was used. The cost function used in the dynamic programming algorithm has various costs: continuity cost, duration cost, pitch cost and energy cost. The continuity cost is the major component of this cost function. It is calculated by comparing the spectral vectors (**Mel binned log FFT vectors**) of two segments at the boundaries. Actually, this cost was computed by comparing the two loudest regions of those two vectors and it was found to be better than a simple Euclidean distance between cepstral vectors.

The dynamic programming algorithm was embedded with a couple of techniques, cost capping and backing off, to ensure high quality synthetic speech. In cost capping, each cost (except continuity cost) is capped at the value which corresponds to the

approximate limit of acceptable signal processing modification. The required specifications are changed to the values corresponding to the capping cost after the segments are selected. A backing off mechanism, similar to one used in decision tree based speech recognition systems, was adopted to correct the acoustic leaf sequences.

Once the best segment sequence is chosen, an energy discontinuity smoothing was applied on it to avoid sudden amplitude changes. Finally, the TD-PSOLA algorithm (Charpentier & Stella 1986, Hamon et al. 1989, Moulines & Charpentier 1990) was used to concatenate the segments and modify pitch, duration and energy values.

1.2.1.7 RSLab

The Lernout & Hauspie RealSpeak Laboratory (RSLab) TTS system is also a corpus-based speech synthesis system (Coorman et al. 2000). This system comprises many components which deal with linguistic processing, prosody prediction, segment selection, concatenation and modification respectively. RSLab uses diphones as units for selection. Two cost functions, unit cost and concatenation cost, are used in unit selection. The unit cost is a multi-dimensional cost function computed on the basis of a combination of symbolic and numerical features. The concatenation cost is also a multi-dimensional cost function and most features used in this computation are directly related to some acoustical property of the candidate units.

The dynamic programming algorithm finds the lowest cost path through all the possible sequences of candidate units by taking account of both of the above costs. In order to avoid the selection of paths which contain one high cost in combination with many low costs (otherwise an utterance with one very bad unit or join will be produced), masking functions were designed. These functions operate on each feature distance measure using two thresholds, the **transparency threshold** and the **quality threshold**. The former controls perception of any discontinuities in pitch, duration and other acoustic features, while the latter takes care of any undesirable speech artifacts. The transparency threshold can be determined by off-line perceptual experiments and quality threshold also can be determined by conducting listening tests.

Also, proper weighting was done to guarantee the relative importance of each feature to the cost being calculated. RSLab did not use any training methods to choose these

weights. Instead they were fine-tuned iteratively by using listening tests. This system also employs additional linguistic knowledge, such as pre-pausal lengthening and /r/ colouring, to achieve context dependency.

1.2.1.8 Summary

Present unit selection systems can be divided into two categories based on whether they use a decision tree-based approach. The advantages and disadvantages of decision tree and non-decision tree-based approaches to unit selection were discussed in (Donovan 2003). Most of the above systems incorporate different techniques such as backing off (Donovan & Eide 1998) or use of masking functions (Coorman et al. 2000) in dynamic programming search to avoid uncommon behaviour of the unit selection and thus ensure high quality synthesis. These techniques are developed to suit their respective systems. It is very difficult to compare the performance of these systems and there have been a few attempts made to evaluate multiple synthesis systems.

In the third ISCA[2] TTS workshop, several TTS systems competed in a comparison test (van Santen, Pols, Abe, Kahn, Keller & Vonwiller 1998). The important factors were: voice quality, number of words correct in semantically unpredictable sentences and overall impression. But there was no formal identification of the best system due to competition rules. An evaluation (Alvarez & Huckvale 2002) of six English TTS systems based on ITU-T P.85 standard was conducted by Alvarez & Huckvale. The main aim of this study was to check the reliability of these tests for the evaluation of TTS systems. The subjective criterion were: overall impression, listening effort, comprehension problems, articulation, pronunciation and voice pleasantness. In this evaluation, the AT&T Next-Gen system (Beutnagel et al. 1999) was ranked first, followed by Speechworks Speechify, RealSpeak (Coorman et al. 2000).

1.2.1.9 rVoice

In this context, we can say a few words about the TTS system used for our research. rVoice is a TTS engine from Rhetorical Systems Ltd (Rutten, Aylett, Fackrell & Taylor

[2]International Speech Communication Association

2002). This uses a state-of-the-art unit selection framework, i.e. finding the optimal unit selection by a Viterbi search for the lowest cost path through the lattice of the target cost and join costs. It is a commercial system, so we can not publish real internal details.

rVoice is a relatively new development[3] and it delivers natural sounding synthetic speech. Internal trials by Rhetorical Systems Ltd. show it to be a very high quality synthesis system. More details of this system are presented in Appendix A.

1.2.2 Join cost functions and spectral distance measures

Since the framework for unit selection synthesis presented in (Hunt & Black 1996) as a search for a lowest cost candidate unit sequence, many join cost functions have been proposed. These functions are mainly derived from various distance measures on speech parameterisations. Usually, these join cost functions are evaluated by conducting some form of perceptual listening experiments.

1.2.2.1 Spectral distance measures based on perceptual studies

Over recent years various researchers have conducted perceptual studies and proposed new spectral distance measures to predict audible discontinuities in concatenative speech synthesis systems. Some of these studies with corresponding investigations are reported below in chronological order.

Klabbers and Veldhuis examined various distance measures on five Dutch vowels to find which measure best predicts concatenation discontinuities (Klabbers & Veldhuis 1998, Klabbers & Veldhuis 2001). These various spectral distances were correlated with the results of a listening experiment, in which listeners had to make a choice between 0 or 1 based on whether the concatenation was smooth (0) or discontinuous (1). They found that a **Kullback-Leibler measure on LPC power-normalised spectra** predicts better than other spectral distance measures: Euclidean distance between formants (F1 and F2), MFCC, the Likelihood Ratio (LR) and the mean squared log-spectral distance (MS LSD) and loudness difference (LD) and expectation differences(ED).

[3]first version released in 2001

A similar study by Wouters & Macon (1998) showed that the **Euclidean distance on Mel-scale LPC-based cepstral parameters** is a good predictor of perceived discontinuity by evaluating several distance measures using perceptual data obtained from listening tests. In these tests, pairs of synthetic mono-syllabic English words that are identical except for one half of a phone, were presented to listeners to rate the difference between the word pairs on a five-point scale. The substituted half phones were limited to three specific cases of vowels. They computed the correlation between objective distance measures and mean listener scores. Their results indicated that distance measures that use a non-linear frequency scale (such as Mel and Bark scales) performed better than those do not. They also found that weighting individual parameters of cepstra, LSF or utilising delta coefficients could improve correlations.

Several acoustic parameters (LPC, linear prediction cepstral coefficients, LSF, MFCC, residual MFCC, bispectrum, modified Mellin transform of the log spectrum, segmental modified Mellin transform of the log spectrum and Wigner-Ville distribution-based cepstrum) were investigated in (Chen & Campbell 1999), for use in assessment and evaluation of synthetic speech. The speech material was synthesised using the CHATR speech synthesis system (Campbell 1996). They first segmented the original speech signal and the synthetic speech signal into frames, each frame represented by several feature coefficients. Then they used dynamic time warping (DTW) for aligning synthetic and natural segments. The overall distortion obtained from the DTW was used as a distance between the synthetic speech and natural speech. Finally, they correlated the distances computed from various acoustic transforms with listener ratings obtained from a mean opinion score (MOS) evaluation. Their results showed that the distances computed using the **bispectrum** had the highest degree of correlation with the MOS scores.

A psychoacoustic experiment on listeners' ability to detect signal discontinuities in concatenative speech synthesis was conducted by Stylianou & Syrdal (2001). They used an experimental version of AT&T Next-Gen (Beutnagel et al. 1999) to synthesise the test stimuli. In this study, the concatenative costs derived from various objective distance measures were compared with listeners' detection scores. These distances were evaluated based on the detection rate, the Bhattacharya measure of separability of two distribu-

tions, and Receiver Operating Characteristics (ROC) curves. Their results showed that a **symmetrical Kullback-Leibler (KL) distance between FFT-based power spectra** and the **Euclidean distance between MFCC** have the highest prediction rates. In contrast to Klabber's results (Klabbers & Veldhuis 1998), this study found that KL distance based on LPC spectra was one of the worst performers.

Another study by Syrdal (2001) focused on phonetic analysis of concatenation discontinuity detection in synthesised speech using either male or female voice. She observed that these discontinuities were significantly more audible for the female voice than the male voice. However, there were many similarities between the results for the two genders. She also observed a higher rate of detection of discontinuities for diphthongs than for monophthong vowels. This is intuitively understandable, since diphthongs have a spectrally dynamic point in the middle, which is usually a joining point.

Donovan (2001) proposed a new spectral distance measure for costing the spectral discontinuities in concatenative speech synthesis. This measure can be described as a decision tree-based context-dependent **Mahalanobis distance between perceptual cepstral parameters**. He conducted listening tests to compare the performance of this new method with other join costs derived from Itakura and KL distances on Mel-binned power spectral; Euclidean and Mahalanobis distances on cepstra, perceptually modified MFCC (P-Cep), log energy and the first and second time differentials of cepstra and P-Cep. The test stimuli were synthesised in a male voice using a modified form of the IBM trainable speech synthesis system (Donovan & Eide 1998). The correlation results showed that this new measure out-performs other measures. Also, further listening tests have justified the use of this measure in the IBM synthesis system.

A comparison of various features – auditory-based(AIM) features (Patterson, Allerhand & Giguere 1995), LPC and MFCC – was presented in (Tsuzaki & Kawai 2002), in order to compute the distances which aid in optimal unit selection. They conducted a perceptual experiment to distinguish synthetic and natural stimuli. The synthetic stimuli were constructed by concatenating units selected according to the possible join cost functions based on the above features. Their results showed that AIM-based join cost functions have a significant advantage over LPCs and similar performance to those using MFCCs.

Using the join cost function to enrich the inventory Klabbers and Veldhuis studied the feasibility of extending the diphone database with context-sensitive diphones, to reduce the occurrence of audible discontinuities (Klabbers, Veldhuis & Koppen 2000, Klabbers & Veldhuis 2001). In order to reduce the number of additional diphones, they used their best spectral distance measure – a **symmetrical Kullback-Leibler distance (SKL) on LPC power-normalised spectra**– in a join cost function to cluster the consonantal contexts that had the same spectral effects on neighbouring vowels. To evaluate the improvements with this additional inventory, they conducted another perceptual experiment. They observed that these additional diphones significantly reduced the amount of audible discontinuities.

A method for enhancing the quality of synthetic speech for the Greek language by reducing the spectral mismatches between concatenated segments was recommended by Founda et al (Founda, Tambouratzis, Chalamandaris & Carayannis 2001). First, they conducted a listening test to determine which spectral distance measure best predicted the audible discontinuities in synthetic speech. They found that a **Kullback-Leibler (KL) distance on power-normalised spectra** was the best predictor among other distance measures they tested. Then, they used this distance measure to enrich the diphone database by providing multiple instances for diphones with large distances. The objective and the subjective evaluation of this procedure indicates a significant reduction in spectral mismatches, thus resulting in better synthetic speech quality.

Summary If there is a single conclusion that can be drawn from the above results, it is that no single join cost function was found to be best in all studies! It is not clear whether this is because the experimental materials vary (small sets of vowels in isolated words, for example), or because join cost is language, accent, or speaker dependent. The latter would have serious implications for synthesisers with numerous languages and voices, such as *rVoice*. Another conclusion that we can draw is that the use of speech parameterisations which include a perceptually-motivated non-linear frequency scale is generally a good idea – a finding consistent with the types of parameterisation used in automatic speech recognition (Huang et al. 2001, Rabiner & Juang 1993).

1.2.2.2 Combined join cost and target cost functions

It is well known that the target cost function determines how well a unit's phonetic contexts and prosodic characteristics match with those required in the synthetic phone sequence. Thus, usually the target cost is computed as the weighted sum of the differences between prosodic and phonetic parameters of target and candidate units. However, attaining the balance between the target cost and the join cost is not easy (i.e. if we give more emphasis on join cost, then target cost may be weighted low and thus result in bad synthesis). One way of lessening this behaviour is to combine these two costs. Here, we present some studies in which these two costs are combined and individual components are weighted based on perceptual/MOS experiments.

1.2.2.3 Functions composed of weighted sub-costs

Chu & Peng (2001) presented a concatenative cost function (weighted sum of several component-costs) as an objective measure for naturalness of speech produced by a Mandarin TTS system. This cost function has seven component-costs: position in phrase, position in word, left phonetic context, right phonetic context, left tone context, right tone context and smoothness cost. A formal mean opinion score (MOS) experiment was conducted in order to know the performance of this concatenative cost function. They optimised weights for the components of this cost function using mean opinion scores and achieved high correlation (-0.872) between objective and subjective measures.

Their recent work with Zhao describes a method for optimising the above cost function (Peng, Zhao & Chu 2002). They used contextual information for all units used in the MOS evaluation obtained in the previous experiment to re-calculate the concatenative cost with a new cost function. Then, they performed optimisation in three steps. First, they optimised the cost function parameters individually. In the second step, interactions among these parameters were explored and they added higher order components to the cost function. Finally, weights were optimised for each component of the new cost function. To validate this new cost function, the correlation between concatenation cost and MOS was used. They achieved a significant improvement in correlation (-0.822 to -0.897).

Minkyu Lee proposed a new method for unit selection in a large corpus-based concatenative synthesis based on a perceptual preference test (Lee 2001, Lee, Lopresti & Olive 2001). His algorithm searches a set of weights for components of the cost function that can produce rankings of renditions that are close to perceptual test results. The downhill simplex method (Press, Teukolsky, Vetterling & Flannery 1992) was used for this multi-dimensional search. A dissimilarity measure was employed to evaluate the closeness of two rankings, obtained from the perceptual test and his method. He found that the unit selection algorithm using the optimised weights chose the same rendition as human listeners preferred in about 83% of the cases (5 out of 6 words).

Phonetic features Blouin, Rosec, Bagshaw & d'Alessandro (2002) presented a concatenation cost based on phonetic and prosodic features to use in unit selection. This cost function is defined as a linear function of weighted sub-costs, computed on various symbolic and prosodic parameters. They optimised weights on these sub-costs by using a multiple linear regression as a function of an acoustic measure of concatenation quality. This acoustic measure is calculated as a *Kullback-Leibler distance on normalised power spectra*. Perceptual evaluation results indicated that the concatenation sub-cost weights determined automatically were better than hand-tuned weights, with or without applying $F0$ and energy smoothing after unit concatenation.

A study by Kawai & Tsuzaki (2002) compared acoustic measures and phonetic features in their ability to predict the audible discontinuities in concatenative speech synthesis. The acoustic measures was derived from MFCCs, mainly Euclidean distances between MFCCs of certain frames. A perceptual listening experiment was used to measure the degradation in naturalness due to signal discontinuities. Then, models were built to predict the degradation scores from acoustic measures and phonetic features. The models used were: multiple regression model; decision tree; and neural network respectively. The multiple regression coefficients were calculated under open and closed conditions of modelling and for acoustic measures and/or phonetic features. Their investigation indicated that phonetic features were more efficient than acoustic measures in predicting the audible discontinuities.

Summary The above studies mainly address the issue of weighing individual components of the join cost function by conducting some perceptual experiments, except one study, in which the weights were optimised based on an acoustic measure of quality (Blouin et al. 2002). One conclusion we can draw here is that either we can use a direct acoustic measure or a measure based on phonetic and prosodic features which correlates well with human perception.

1.2.3 Linear dynamic models for speech modelling

Linear dynamic systems, also known as Kalman filter models, have already been employed in speech research in recent years for acoustic signal modelling, especially in speech recognition as an alternative to frame-based models such as HMMs (Digilakis, Rohlicek & Ostendorf 1993, Frankel & King 2001, Frankel 2003). These models characterise the spatial correlations of observed data and also model the temporal characteristics with their dynamic nature. Here, we survey a number of papers on linear dynamic models in the context of speech modelling.

Digilakis et al. (1993) proposed an approach based on the Expectation-Maximisation (EM) algorithm for training dynamical-system segment models. Their classification experiments showed significant improvement of this model over models using either an independent-frame or a Gauss-Markov assumption using the observed frames in a segment of speech. However, the dynamical system model is computationally more expensive than those frame-based models.

A method based on Kalman filtering for obtaining smooth vocal tract parameters from the analysis during the close phase of the glottis was proposed in (McKenna & Isard 1999, McKenna 2001). This method treats speech data during excitation and open phases as **missing data** to avoid adverse effects on vocal tract filter parameters. The predictor coefficient trajectories produced with this method are far smoother than those obtained from closed-phase covariance analysis. Also, this approach automatically determines the closed phase location from speech signal itself. Finally, they observed that this method overcomes flaws in traditional fixed frame linear predictive analysis, and offers accurate separation of source and filter.

Smith, de Frietas, Robinson & Niranjan (1999) presented a method for formant tracking using LDMs, mainly comparing various initialisation schemes for the EM algorithm. They modelled each frame of speech waveform with a linear time-invariant stochastic state space model. The model parameters were computed using an EM algorithm. In their experiments, three schemes, 4SID (Subspace State Space System Identification), ARMA (Auto-Regressive Moving Average) and AR(1) were compared to initialise EM parameters. Formant trajectories were estimated using EM with these three initialisation schemes and 4SID-EM produced the best formant tracks and higher likelihoods.

More recently, this model has been applied to speech recognition by Frankel (2003) in his PhD research. He compared the performance of systems which model acoustic, articulatory and both acoustic-articulatory features using speaker-dependent database from MOCHA corpus (Wrench 2001). A stack decoder with A^* search was used for more efficient decoding for segment models.

First, classification experiments were carried out using both real and automatically recovered articulatory features. The automatically estimated articulatory features were generated using neural networks (Richmond 2002). Acoustic features used in these experiments were MFCC, PLP (Perceptual Linear Prediction) cepstra. The best classification results were obtained with combined real articulatory and acoustic (MFCC & energy) features. Further, experiments on the same data confirmed that combined acoustic and real articulatory features resulted in high recognition accuracy.

1.3 Objective of this thesis

Many recent commercial and research synthesisers (Black & Campbell 1995, Huang et al. 1996, Donovan & Eide 1998, Beutnagel et al. 1999, Coorman et al. 2000), show that the unit selection method performance better than diphone synthesis. But still, there are many concerns in selecting optimal speech segments/units from the database, preparing the speech database and in producing high naturalness in synthetic speech. Also, the unit selection technique is renowned for both high quality synthesis output and occasional extremely low quality output (Black & Taylor 1997*a*). Hence, the main

issue of current research is to minimise the number of these bad synthesis examples.

The above issues relate to the formulation of the two cost functions, namely **target cost** and **concatenation (join) cost**, as these are used in selecting for the optimal units from the database thus resulting in high quality natural sounding synthetic speech. Past experience (Klabbers & Veldhuis 1998, Chu & Peng 2001, Stylianou & Syrdal 2001, Donovan 2001) shows that the performance of existing join cost functions in representing human perception to concatenation discontinuities is not very satisfactory. It is also known that the join cost plays an important role in unit selection with respect to computational requirements in addition to producing high quality natural speech.

Unfortunately, there is no standard measure to evaluate the join cost function. This is a hard problem in current state-of-the-art unit selection concatenative speech synthesis research (van Santen, Wouters & Kain 2002). Recently, various researchers have come up with several objective measures to evaluate this join cost. However, there is no agreement among them. In addition, these perceptual studies (reported in the literature review, section 1.2) have focused on human detection of audible discontinuities in **isolated words** generated by concatenative synthesisers. Hence, it is worthwhile carrying out such experiments on whole sentences to incorporate contextual effects.

All these factors necessitate a proper study of how various objective distance measures correlate with human perception to audible discontinuities in whole sentences. In addition, we would like to know which speech parameterisation could be used to compute these objective distance measures. Previous studies mainly concentrated on linear prediction coding (LPC) based parameters and cepstral coefficients as acoustic vectors for computing distances. Previous research on rule-based systems (Allen et al. 1987) suggests that formant-like parameters are a better representation of co-articulation and other spectral variations of speech. A set of formant-like features, called **Multiple centroid analysis (MCA)** parameters, proposed by Crowe & Jack (1987), might be well suited for this purpose.

In the first part of this thesis, we investigate several objective distance measures based on Mel-frequency cepstral coefficients (MFCC), line spectral frequencies (LSF), and MCA coefficients by conducting a perceptual test. This test uses synthetic examples

of **whole (continuous) sentences with polysyllabic words** generated using an experimental version of the state-of-the-art commercial unit selection synthesis system, *rVoice* from Rhetorical Systems Ltd[4].

Once appropriate units are concatenated, most systems attempt some form of local parameter smoothing to disguise the remaining discontinuities. The join cost measure and the join smoothing method interact closely. If we have a sufficiently large database and a perfect join cost measure then no smoothing would be required. Conversely, if we could smooth joins better, then the method of computing join cost would be less critical and a smaller inventory might be adequate for synthesis. This suggests that combining the computation of join cost with the operation of smoothing, example – by using a common representation and/or a common model, should give improved results.

The second part of this thesis concentrates on combining join cost computation and join smoothing by learning an underlying representation from the acoustic signal. In order to accomplish this task, we have chosen linear dynamic models (LDMs), sometimes known as Kalman filters. Once these models are trained, the join cost is computed based on the error between model predictions and actual observations (acoustic parameters). These models also smooth actual, noisy observations; which are then can be used to synthesise speech.

Another listening test is designed to compare three join cost functions and different smoothing methods in the third part of this thesis. The three join cost functions are chosen based on their correlations with mean listener scores obtained from our first perceptual test. The smoothing methods are: no smoothing, linear smoothing and smoothing obtained using a Kalman filter.

1.4 Thesis structure

Chapter 2 provides a detailed description of join cost computation, concentrating on the spectral component of the join cost function. We briefly describe various speech parameterisations and distances used in our study. In Chapter 3, we discuss our first perceptual listening test design and present correlations of listener scores with our var-

[4]www.rhetorical.com

ious objective spectral distance measures. Chapter 4 presents correlation results with weighted spectral distances and distances computed on various combinations of speech parameters.

We introduce linear dynamic models in Chapter 5 and discuss join cost computation using these models. In Chapter 6 we evaluate three join cost functions and different smoothing techniques by conducting a listening test. Finally in Chapter 7 we provide conclusions of this thesis.

1.5 Publications

Some of the material contained in this thesis previously appeared as papers in conference proceedings: Vepa, King & Taylor (2002*a*), Vepa, King & Taylor (2002*b*) and Vepa & King (2003). A forthcoming book chapter (Vepa & King forthcoming) and a conference paper (Vepa & King 2004) are also based on the material of this thesis.

Chapter 2

Join Cost

2.1 Introduction

As we have already seen in section 1.1.3, unit selection speech synthesis systems use two cost functions, namely **target cost** and **join cost**, for selecting appropriate units from a large database.

The target cost reflects how well a candidate unit's prosodic and phonetic properties and contexts match those required in the synthetic phone sequence. Hence, this is usually calculated as weighted sum of differences between prosodic and phonetic parameters of target unit and candidate unit. The join cost computation is intended to achieve smooth concatenation between neighbouring units, to minimise audible discontinuities. Thus the join cost function must measure how well two successive speech units join together.

Though these two costs are equally important for optimal unit selection, we dedicate one chapter to join cost since the main objective of this thesis is to design and evaluate join cost functions. Hence, it is worthwhile to discuss how the join cost is computed, what are the various sub-components and which acoustic features and distance metrics could be used, for clear understanding of later chapters to the reader.

In this chapter, we discuss join cost in detail with more focus on the spectral cost component of the join cost function. In section 2.2.1, we discuss various sub-costs contributing to join cost. Then in section 2.3, we describe various speech parameterisations: Mel Frequency Cepstral Coefficients (MFCC), Line Spectral Frequencies (LSF) and Mul-

tiple Centroid Analysis (MCA) coefficients, used in our study. Finally, various distance measures are described in section 2.4.

2.2 Join cost computation

The ideal join cost is one that correlates highly with human perception of discontinuity at unit concatenation points. In other words, it should predict the degree of perceived discontinuity. Though researchers have come up with many join cost functions based on the spectral parameters of the speech units being concatenated to compute join cost, it is very difficult to formulate some objective method to measure this discontinuity, mainly due to co-articulation phenomena. As these phenomena affect the spectral envelope of diphthong, vowel and vowel-like sounds drastically (most consonants do not undergo drastic changes due to co-articulation), the spectrum of the same diphthong or vowel will appear differently in different contexts; for example the spectrum of the vowel *eh* in words "get" and "bet".

2.2.1 Join cost function

Usually, a join cost function is derived from a few sub costs defined on spectral and prosodic properties of the speech frames on either side of a proposed join between two candidate units. For example :

- *difference in amplitude (energy)*
- *difference in fundamental frequency ($F0$) (perceived as pitch)*
- *mismatch in spectral envelope (parameters)*

Since energy and fundamental frequency ($F0$) are scalars, i.e. one value for each frame of speech on either side of the join, a simple absolute difference (distance) between the energies of two candidate units could be used (Hunt & Black 1996). Usually, a logarithmic scale is used to reduce the dynamic range of energy and also to mimic the human perceptual property of compression. $F0$ can be represented on various perceptual scales, such as Mel, Bark or semi-tone.

The spectral parameterisation is often a smoothed spectral envelope, i.e. an estimate of the vocal tract frequency response, which is derived from a short-term analysis of the speech signal. Examples of such parameterisations are: Linear Prediction (LP) spectrum (Makhoul 1973, Makhoul 1975); Mel Frequency Cepstral Coefficients (MFCC) (Rabiner & Juang 1993, Huang et al. 2001); Line Spectral Frequencies (LSF) (Itakura 1975); Multiple Centroid Analysis (MCA) coefficients (Crowe & Jack 1987); Perceptual Linear Prediction (PLP) spectrum (Hermansky 1990).

Usually, the computation of these acoustic feature vectors (for F0 and energy there is only one value in the feature vector) is done off-line to reduce the computational complexity of the system at run time. This requires storage of the feature vectors – so compact representations are preferable. It is a normal practice to store the initial and final frames of each speech unit in the database for distance calculations. Then, distances between frames at either end of concatenated units will be computed at run time. More complicated methods may use multiple frames of the both candidate units but the price one has to pay is in terms of computational complexity and memory requirements. Also, one can compute the costs offline and cache them. But, this requires very large storage space, due to the millions of possible joins for a large database.

It is rather easy to compute $F0$ and energy distances (indeed, costs[1]) compared to spectral envelope mismatches (spectral distances). For $F0$, it is mainly representing them in some perceptual scale, e.g. in Bark or Mel, and then obtaining the difference between them. The case for energy is similar: a logarithmic compression, such as decibels, could be used. However, for spectral envelope, it is a hard problem for the following reasons:

- *Many spectral representations (parameterisations) are available in the literature.*

- *Many statistical distances are applicable, e.g. Euclidean, Mahalanobis, Kullback-Leibler, and other popular distances used in speech processing (Gray & Markel 1976), such as Itakura-Saito and COSH.*

[1] We refer **cost** as the value obtained by multiplying distances with proper weights on each of these three components, $F0$, energy and spectral parameters.

In later chapters of this thesis, many combinations of several spectral parameterisations and distances are investigated to achieve better join cost functions. With these many combinations, it is already a complex problem and if we include $F0$ and energy components it will be cumbersome. Hence, in this thesis we concentrate on the **spectral component of the join cost function**.

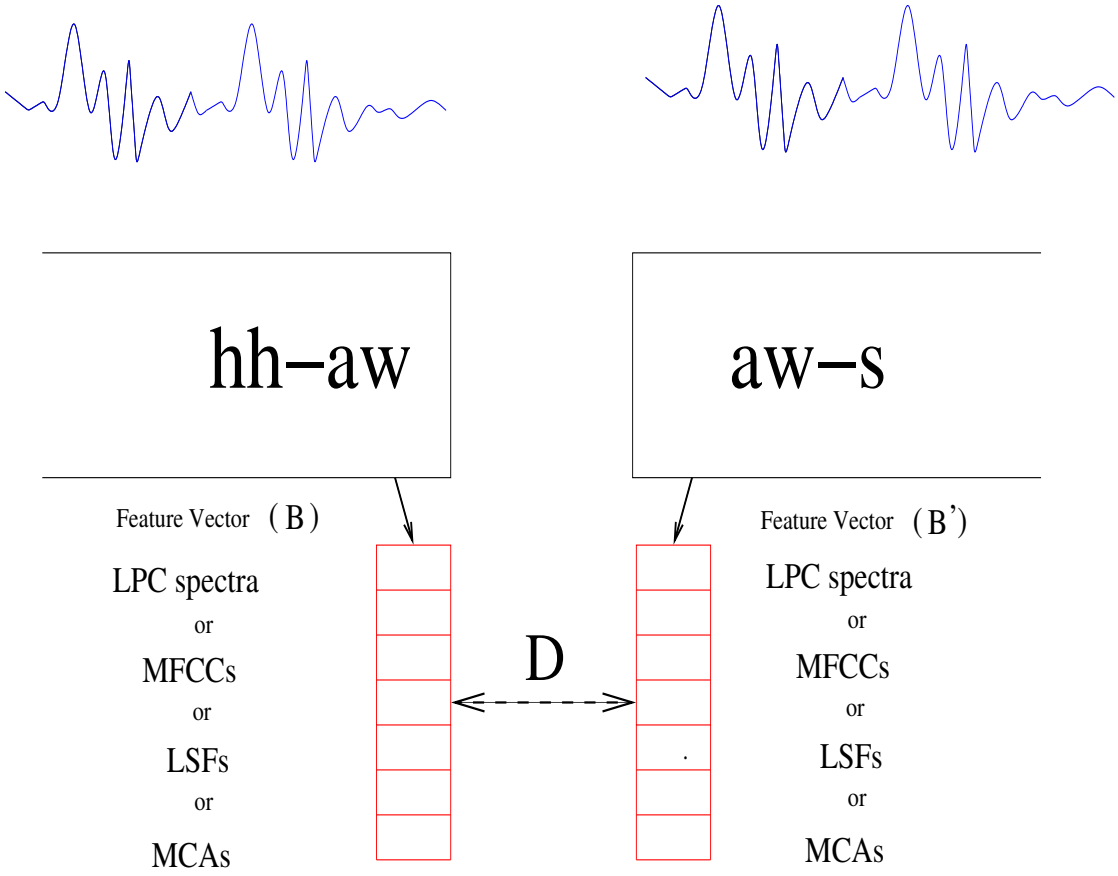

Figure 2.1: Spectral distances

The spectral distance between corresponding *speech parameterisations* of the final and initial frames of two units to be concatenated is described in figure 2.1. We used fixed-frame rate (as opposed to pitch-synchronous frames) as Rhetorical Systems Ltd. provided us access to these files only for the male voice (refer to Appendix A for details of voices used in this research). Consider two diphones, *hh-aw* and *aw-s*, which form a target unit sequence. To compute the spectral distance between candidates for these two diphones, feature vectors of the final frame of diphone *hh-aw* and the initial frame

of diphone *aw-s* have to be obtained. As we have mentioned already, these feature vectors are usually computed off-line and stored in the database. Once these features are available, various distance measures can be applied to them.

2.3 Speech parameterisations

Four different speech parameterisations: LPC spectra, MFCCs, LSFs and MCA coefficients, were used in our study. The computation aspects and other characteristics of each of them (except LPC, which is briefly discussed in section 6.2.3) are discussed in the following sections.

2.3.1 Mel-frequency cepstral coefficients

Mel-frequency cepstral coefficients (MFCCs) (Davis & Mermelstein 1980) are perceptually motivated speech parameters, which approximate the behaviour of the human auditory system and also crudely separate source and filter (explained in 2.3.1.1). Because of its non-linear frequency scale, decorrelated nature and robustness to noise, this representation is widely used in speech recognition. Since speech recognition research influences synthesis in many respects, this representation has also been inherited by speech synthesis.

2.3.1.1 MFCC computation

Usually, MFCC parameters are computed from the discrete Fourier transform (DFT) of the windowed speech signal using Mel-binned filters. $X[k]$ is the DFT of one frame of the speech signal, $x[n]$:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{\frac{-j2\pi nk}{N}} \qquad 0 \leq k < N \tag{2.1}$$

Then, a filter bank with triangular filters spaced along a Mel-warped frequency scale is constructed, as shown in figure 2.2. This filter bank is applied to the DFT coefficients. The outputs of the filter bank are then subjected to logarithmic compression. This logarithmic operation is a *homomorphic* transformation, a transform that converts a convolution into a sum. Finally, a discrete cosine transform (DCT) is used to compute

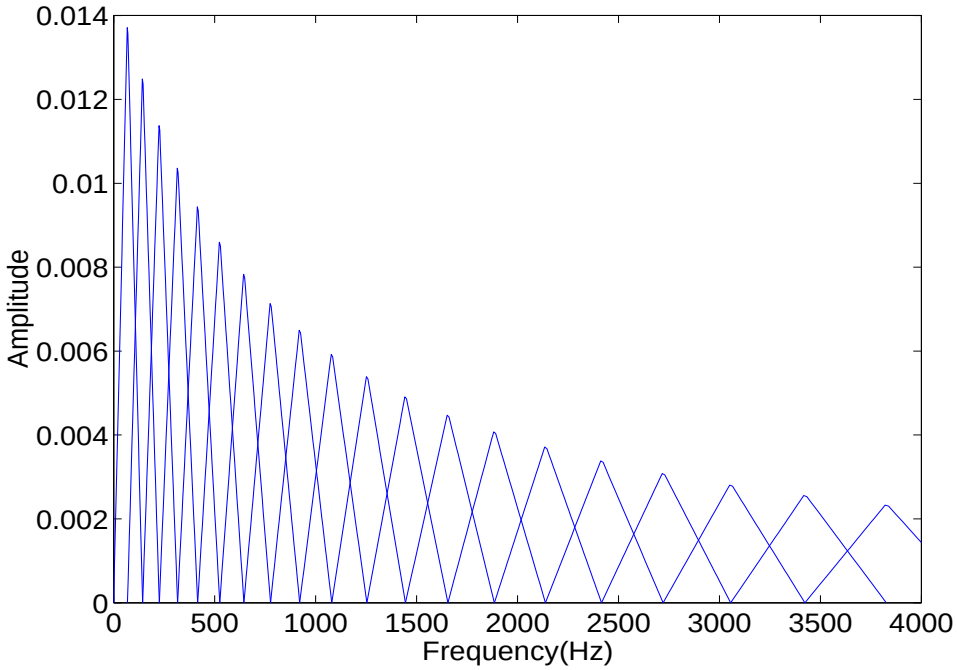

Figure 2.2: Mel-warped triangular filters

the MFCCs from the above log Mel-binned filter bank coefficients (m_j), as shown below (Young, Odell, Ollason, Valtchev & Woodland 1997):

$$MFCC_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \; cos\left(\frac{\pi i}{N}(j - 0.5)\right) \tag{2.2}$$

Figure 2.3 shows the first and second MFCC parameters of a sample speech signal. Typically, the higher order cepstral coefficients are significantly smaller in magnitude than the lower ones. Hence, in order to normalise the magnitudes of these coefficients, cepstral liftering is used. This is done by using the following formula (Young et al. 1997):

$$c_n^{'} = \left(1 + \frac{L}{2} \; sin\frac{\pi n}{L}\right) c_n \tag{2.3}$$

where L is the lifter parameter ($L = 2(N-1)$), c_n and $c_n^{'}$ are cepstral coefficients before and after liftering.

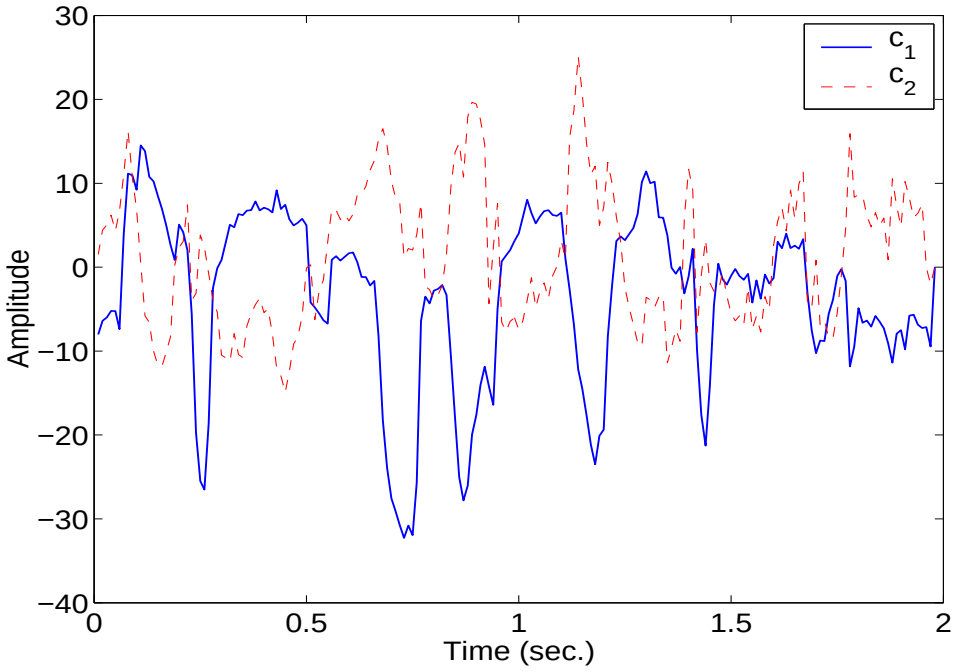

Figure 2.3: First and second MFCCs of a sample speech signal

2.3.2 Line spectral frequencies

The line spectrum pair (LSP) representation of speech was first introduced by Itakura (1975) as an alternative representation of linear predictive coding (LPC) coefficients. The LSPs are the roots of symmetrical and anti-symmetrical polynomials constructed from the LPC polynomial. The angles of these roots are called line-spectral frequencies (LSFs).

These are widely used in speech processing, mainly in speech coding because of their spectral characteristics and statistical properties, including a bounded dynamic range, a sequential ordering of the parameters and a filter stability preservation property. LSFs are also uncorrelated, i.e. the covariance matrix is approximately diagonal (as speech is not quite a stationary autoregressive process), when they are estimated from stationary autoregressive processes (Erkelens & Broersen 1995). Additionally, LSFs/LSPs lend themselves to time interpolations with a relatively low spectral distortion (Umezaki & Itakura 1986). Because of these properties they are also extensively used in speech synthesis.

2.3.2.1 LSF computation

LSFs can be computed from an all-pole model of speech. Consider the LPC (order p) analysis of a speech segment, represented in the Z-domain:

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} \tag{2.4}$$

where a_k are the LPC coefficients. Two artificial ($p+1$ order) polynomials are defined corresponding to a complete opening and a complete closure of the glottis in an acoustic tube model (Soong & Juang 1984), as shown below:

$$P(z) = A(z) - z^{-(p+1)} A(z^{-1}) \tag{2.5}$$

$$Q(z) = A(z) + z^{-(p+1)} A(z^{-1}) \tag{2.6}$$

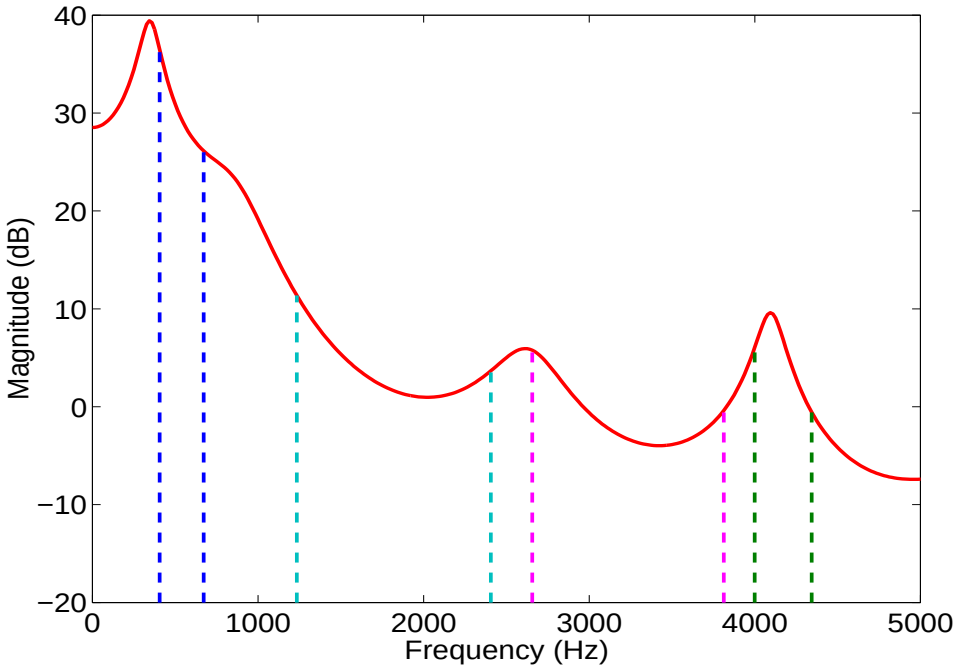

Figure 2.4: LPC spectrum envelope and the first four pairs of LSFs

These symmetric ($P(z)$) and anti-symmetric($Q(z)$) polynomials have the following important properties.

1. All roots of $P(z)$ and $Q(z)$ lie on the unit circle

2. All roots are distinct and interlaced with one another

3. Minimum phase property of $A(z)$ (i.e. all poles and zeros are inside the unit circle) is easily preserved after quantisation of the zeros of $P(z)$ and $Q(z)$

Because of the third property, the stability of the LPC synthesis filter can be easily ensured. When the constant roots at $z = 1$ and $z = -1$ are removed from the above polynomials, the remaining roots are called line spectrum pairs (LSP). The frequencies corresponding to these LSPs are called line spectral frequencies (LSF). Figure 2.4 shows the LSFs overlayed on an LPC spectrogram.

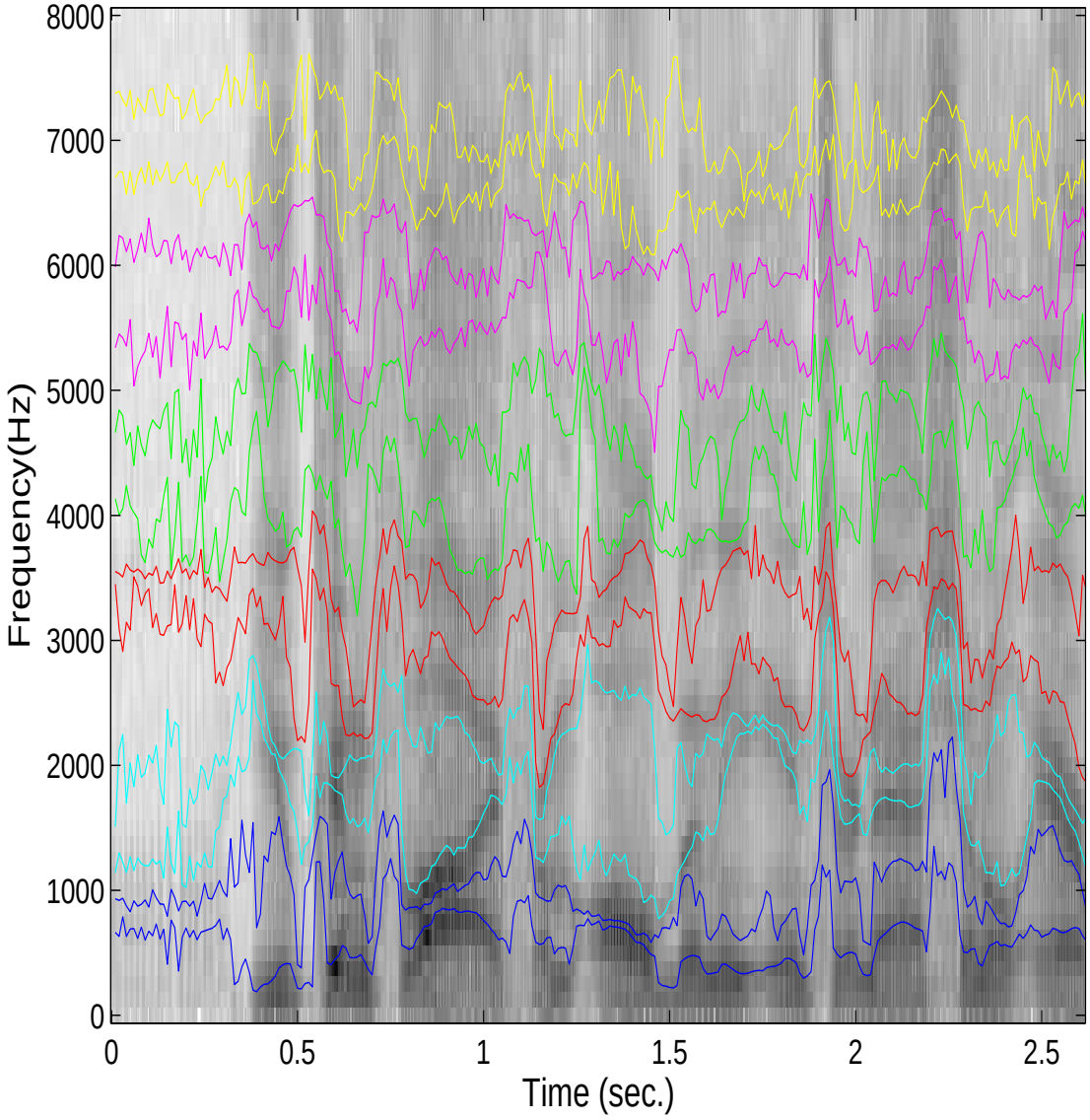

Figure 2.5: LSF trajectories overlayed on spectrogram of a sample speech waveform

2.3.2.2 LSF spectral characteristics

Figure 2.5 shows trajectories of the line spectral frequencies overlayed on the spectrogram of a sample speech signal. If more than one LSF occurs in a certain frequency band, then there is a strong resonance (Sugamura & Itakura 1986). Conversely, LSFs are spaced equally along the frequency axis when there is no resonance. Thus a cluster of 2 or 3 LSFs characterises a formant frequency and the bandwidth depends on the closeness of the corresponding LSFs.

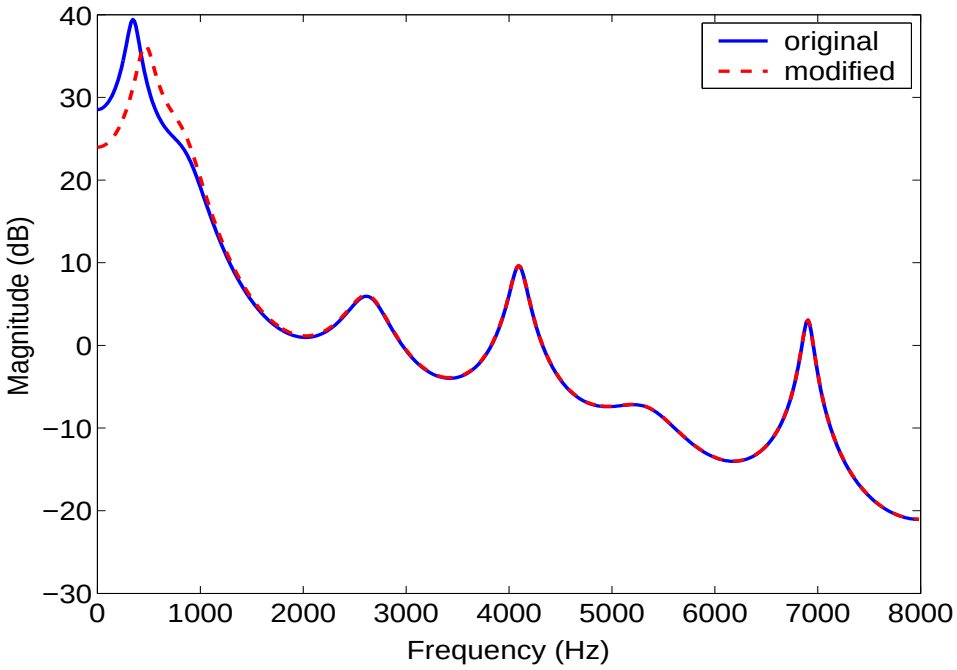

Figure 2.6: The localised spectral sensitivity property, LPC spectrum before and after modifying the first two LSFs

The spectral sensitivities of LSFs are localised. i.e a change in a given LSF produces a change in the LPC spectrum only in its neighbourhood. This is illustrated in figure 2.6, in which first two LSFs are modified (increased by 100 Hz) and the corresponding LPC spectrum is plotted over the original LPC spectrum. In the figure, the spectra differ only in the low frequency portion and remain the same elsewhere. The localised spectral sensitivity property makes LSFs ideal for scalar quantisation as the individual LSFs can be quantised independently (Paliwal & Kleijn 1995).

2.3.3 Multiple centroid analysis

Multiple Centroid Analysis was introduced by Crowe & Jack (1987) as an alternative to formant tracking using traditional peak picking techniques. This algorithm employs a generalised centroid (centre of gravity) estimation to describe a spectral distribution as a fixed number of peaks.

The MCA algorithm does not use an all-pole model, thus it can also be used to parameterise nasals, fricatives or high pitched speech (Wrench, Watson, Soutar, Robertson & Laver 1994). This is more robust than peak picking and does not involve any ambiguity about extra peaks in the spectrum. Because of the above properties, MCA is an attractive alternative to linear prediction-based formant trackers.

2.3.3.1 Centroid computation

To compute centroids, we consider a multi-modal distribution such as a speech power spectrum, then split it into an appropriate number of partitions (say 4 or 5) as shown in figure 2.7. The centroid (k) of a specific partition of the distribution $P(n)$ bounded by

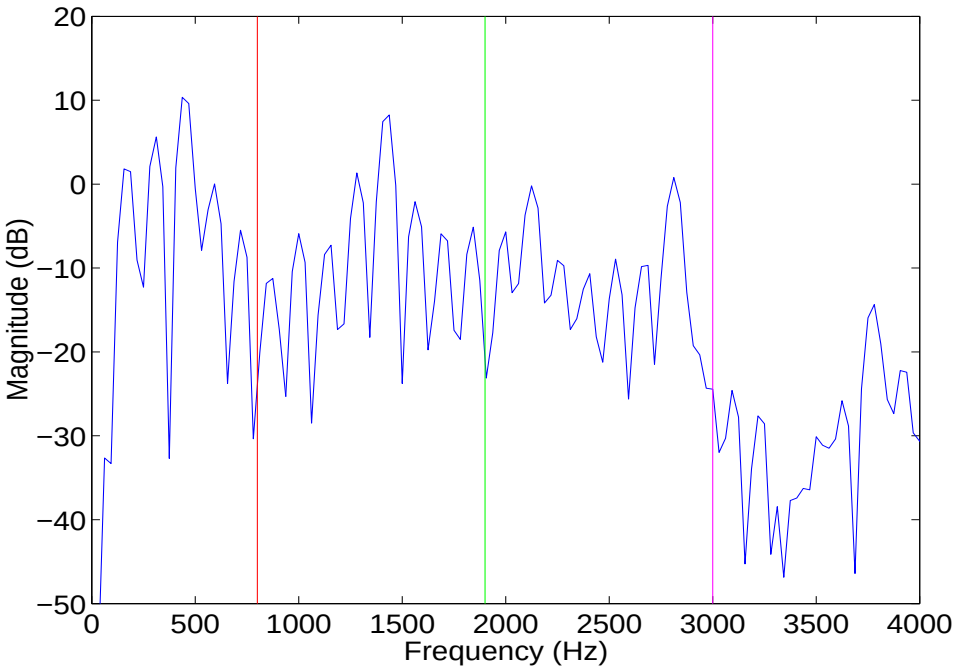

Figure 2.7: Speech power spectrum and MCA (four partitions).

$n = c_1$ and $n = c_2$ (where n is frequency index and $P(n)$ is power spectrum) is estimated as the value that gives minimum squared error, as shown in the equation below:

$$e(c_1, c_2, k) = \sum_{n=c_1}^{c_2} (n-k)^2 P(n) \tag{2.7}$$

This will be computed for every possible combination of partitions and a minimum error condition is used to determine the optimal partition boundary positions. If the spectral distribution within a single partition contains a single formant then the centroid and associated variance represents the formant frequency and bandwidth (Wrench 1995).

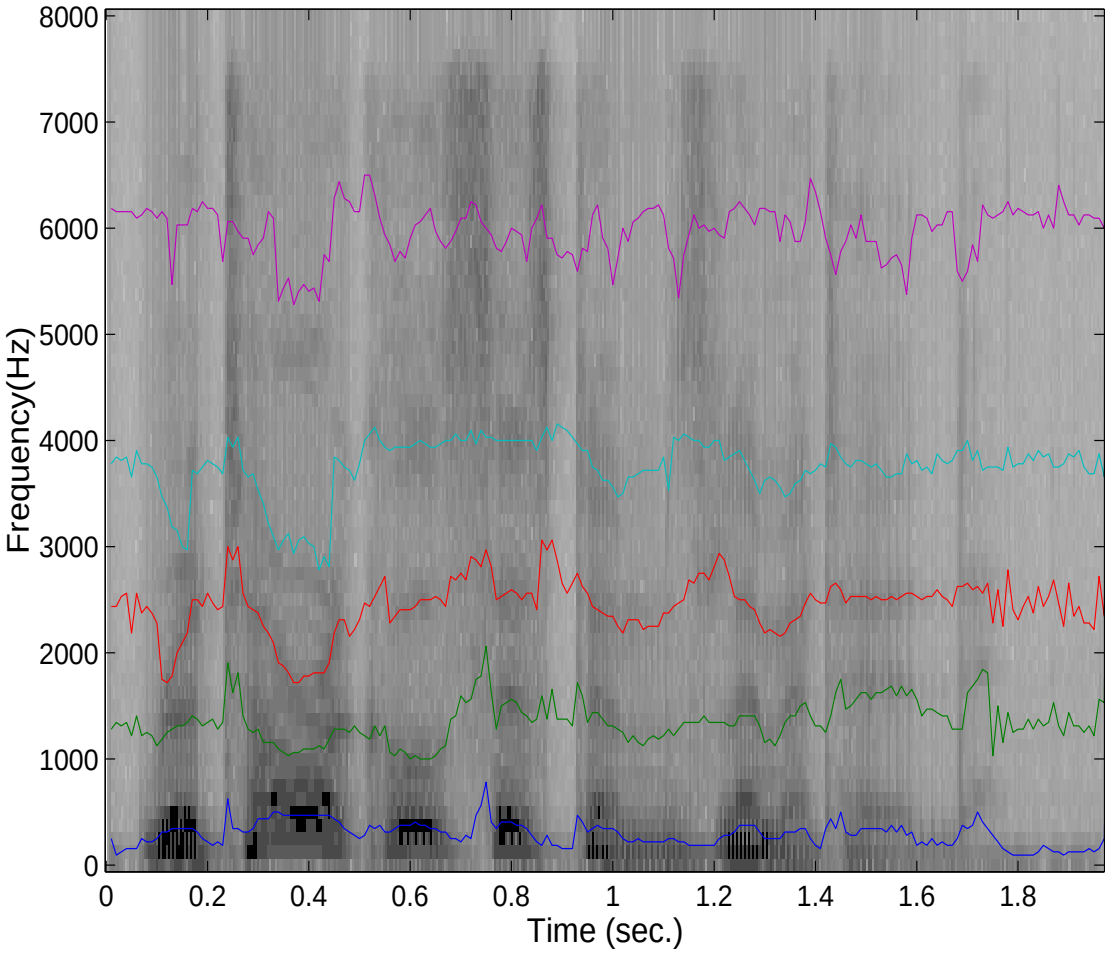

Figure 2.8: MCA (formant) trajectories overlayed on the spectrogram of a speech signal

Figure 2.8 shows five MCA coefficients overlayed on the spectrogram of a speech signal, which are corresponding frequencies of the five centroids of the power spectrum. These can be interpreted as the formants for speech sounds that have them. Usually, the

partitions are constrained to a range of frequencies associated with each formant, to reduce the computational complexity. Also, these constrains make the formant estimation more robust (Wrench et al. 1994).

2.3.4 Delta coefficients

Delta coefficients (Furui 1986) are first order time derivatives of any of the above speech parameters. These are computed in terms of the corresponding static parameters. A simple method of computing deltas is shown below:

$$d[t] = x[t] - x[t-1] \tag{2.8}$$

which is also called a *simple difference*. More sophisticated methods can also be used to compute these delta coefficients, such as the following regression formula used in HTK (Young et al. 1997):

$$d[t] = \frac{\sum_{\theta=1}^{\Theta} \theta(x[t+\theta] - x[t-\theta])}{2\sum_{\theta=1}^{\Theta} \theta^2} \tag{2.9}$$

where d_t is delta coefficient at time t, $x[t+\theta], x[t-\theta]$ are corresponding static coefficients, and typically $\Theta = 2$.

2.4 Distance measures

To measure the difference between two vectors of speech parameters we need a distance measure. This may be a *metric*, provided it has the required properties[2], but this is not necessary. Examples of such measures are: absolute magnitude distance; Euclidean distance; Mahalanobis distance – i.e. Euclidean distance normalised for (co)variance; Kullback-Leibler (KL) divergence. All but the KL divergence are metrics. We have used a symmetrical version of KL divergence to compute the distance between two speech parameterisations, as explained in section 2.4.2.

[2]symmetrical; triangular inequality property

2.4.1 Simple distance measures

2.4.1.1 Absolute distance

Simple absolute distance between two feature vectors, X and Y is computed as the sum of the absolute magnitude difference between individual features of the two feature vectors:

$$d_{Abs}(X,Y) = \sum_{i=1}^{N} |X_i - Y_i| \tag{2.10}$$

2.4.1.2 Euclidean distance

The Euclidean distance is computed as shown below:

$$d_{Eu}(X,Y) = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2} \tag{2.11}$$

This distance is very easy to compute. However, the Euclidean distance does not take any account of variances or co-variances of the distribution of the feature vectors.

2.4.2 Statistically-motivated distance measures

Popular distance measures from the field of statistics include the Kullback-Leibler divergence (which is not a metric) and the Mahalanobis distance metric. We have no prior preference for any of these: we will find out which is best by experiment.

2.4.2.1 Mahalanobis distance

Mahalanobis distance, also used by Donovan (2001) in a join cost function, is a generalisation of standardised (Euclidean) distance in that it takes account of the variance or covariance of individual features. In general, the Mahalanobis distance metric requires the estimation of covariance matrices; often, the off-diagonal elements of the covariance matrix are assumed to be zero – this saves computation and storage. The Mahalanobis distance between two feature vectors X and Y with full covariance matrix (Σ) is shown below:

$$d_{Ma}(X,Y)^2 = (X-Y)^T \, \Sigma^{-1} \, (X-Y) \tag{2.12}$$

In preliminary experiments, we found that making this diagonal covariance assumption was reasonable: using full covariance matrices did not improve performance over using diagonal matrices for MFCCs. The Mahalanobis distance with diagonal covariance matrix is shown in equation 2.13.

$$d_{Ma}(X,Y)^2 = \sum_{i=1}^{n} \left[\frac{X_i - Y_i}{\sigma_i} \right]^2 \tag{2.13}$$

where, σ_i is the standard deviation of the i^{th} feature of the feature vectors, i.e. a diagonal entry of the covariance matrix.

2.4.2.2 Kullback-Leibler (KL) divergence

The Kullback-Leibler (KL) divergence (Kullback & Leibler 1951) computes the "distance" between two probability distributions. It has been used in join cost functions before (Klabbers & Veldhuis 1998, Stylianou & Syrdal 2001, Donovan 2001, Founda et al. 2001), and requires us to interpret spectral envelopes as probability distributions. Let $P(\omega)$ and $Q(\omega)$ denote two spectral envelopes, then KL divergence is defined as:

$$d_{KL}(P,Q) = \frac{1}{2\pi} \int_0^{2\pi} P(\omega)\ log \frac{P(\omega)}{Q(\omega)} d\omega \tag{2.14}$$

The above divergence is asymmetric, so we can define a symmetric KL "distance" (not a metric) as:

$$\begin{aligned} d_{SKL}(P,Q) &= (D_{KL}(P,Q) + D_{KL}(Q,P))/2 \\ &= \frac{1}{4\pi} \int_0^{2\pi} (P(\omega) - Q(\omega))\ log \frac{P(\omega)}{Q(\omega)} d\omega \end{aligned} \tag{2.15}$$

The standard procedure for evaluating the above equation is by performing the integral as a summation over discrete frequencies. However, Veldhuis & Klabbers (2003) recently showed that this approximation is inferior to the exact method for computing symmetrical KL distance, equation 2.15, for all-pole (LPC) spectra. The computational cost of this exact method is substantially higher. We used the discrete summation approximation:

$$d_{SKL}(X,Y) = \sum_{i=1}^{N} (X_i - Y_i)\ log \frac{X_i}{Y_i} \tag{2.16}$$

2.4.2.3 Itakura-Saito (I-S) distance

Itakura and Saito proposed a distortion measure, in their formulation of linear prediction, as an approximate maximum likelihood estimation (Rabiner & Juang 1993) given as:

$$d_{IS}(S,S^{'}) = \int_{-\pi}^{\pi}[e^{V(w)} - V(w) - 1]\frac{dw}{2\pi} \tag{2.17}$$

where, $V(w) = log(S(w)) - log(S^{'}(w))$, $S(w)$ and $S^{'}(w)$ are power spectra. The main assumption in deriving this measure was that the speech was generated by a Gaussian process (Gray & Markel 1976). We have used the discrete version of the I-S distance, as shown below:

$$d_{IS}(X,Y) = \frac{1}{N}\sum_{i=1}^{N}\Big[\frac{X(w_i)}{Y(w_i)} - log\Big(\frac{X(w_i)}{Y(w_i)}\Big) - 1\Big] \tag{2.18}$$

where $X(w_i)$ and $Y(w_i)$ are LPC power spectra of the two frames at the join. The I-S distance is asymmetric; the symmetric version is obtained as the simple average of two asymmetric versions.

$$\begin{aligned} d_{SIS}(S,S^{'}) &= (d_{IS}(S,S^{'}) + d_{IS}(S^{'},S))/2 \\ &= \frac{1}{2}\int_{-\pi}^{\pi}[e^{V(w)} - V(w) - 1 + e^{-V(w)} + V(w) - 1]\frac{dw}{2\pi} \qquad (2.19) \\ &= \int_{-\pi}^{\pi}[\frac{e^{V(w)} + e^{-V(w)}}{2} - 1]\frac{dw}{2\pi} \qquad (2.20) \\ &= d_{COSH}(S,S^{'}) = \int_{-\pi}^{\pi}[coshV(w) - 1]\frac{dw}{2\pi} \qquad (2.21) \end{aligned}$$

Hence, the above distance is also termed a **COSH** distance (Gray & Markel 1976). We have used the discrete version:

$$d_{COSH}(X,Y) = \frac{1}{2N}\sum_{i=1}^{N}\Big[\frac{X(w_i)}{Y(w_i)} + \frac{Y(w_i)}{X(w_i)} - 2\Big] \tag{2.22}$$

Chapter 3

Spectral Distance Measures

3.1 Introduction

Over the years, many distance measures have been applied to speech processing (Gray & Markel 1976) and speech recognition (Rabiner & Juang 1993) to measure the dissimilarities between two spectral vectors. Also, various speech parameterisations, such as FFT power spectra, linear prediction, Mel-cepstra and perceptual linear prediction (Hermansky 1990), have been proposed. There are many combinations of distances and parameterisations to explore and investigate for the purpose of the optimal join cost function in unit selection synthesis. This forms one of the main goals of this thesis.

In speech synthesis, the objective of these spectral distance measures is to predict audible discontinuities in concatenated units and aid in optimal selection of speech units from a large database. In order to achieve this objective, a perceptual evaluation could be the best way of assessing various spectral distance measures. Recently, researchers (Klabbers & Veldhuis 1998, Wouters & Macon 1998, Donovan 2001, Stylianou & Syrdal 2001) have come up with different perceptual study designs to evaluate these distance measures. A detailed discussion of these studies was can be found in the literature review, section 1.2.2. To have continuity for the reader, we summarise the results of these perceptual studies in table 3.1. These studies provide the basis for our perceptual listening experiment, though our design is very different.

In this chapter, we first describe the design and procedure of our perceptual listening

experiment. Also we discuss our methodology to obtain perceptual data from these experiments. In section 3.3, we present correlations between various spectral distance measures and the perceptual data. These distance measures are computed on spectral vectors of single frames on either side of the join. We investigated distances based on multi-frames and computed correlations, which are presented in section 3.3.2. Finally, we discuss the performance of these spectral distance measures. In Chapters 4 & 5, we will look at some more sophisticated measures.

Reference	*Test material*	*Best prediction measure*
(Klabbers & Veldhuis 1998)	2248 CVC words with Dutch vowels: *a, A, i, I, u* (in SAMPA notation)	Kullback-Leibler measure on LPC power-normalised spectra
(Wouters & Macon 1998)	166 English words with vowels: *aa, ae, iy, uw* (in OGI-bet notation)	Euclidean distance on Mel-scale LPC-based cepstral parameters
(Chen & Campbell 1999)	Six Japanese sentences (1-natural, 5-synthetic)	Bi-spectrum
(Stylianou & Syrdal 2001)	Modified rhyme test (MRT) set containing 335 words	Kullback-Leibler measure on FFT-based power spectra
(Donovan 2001)	112 English CV pairs of different phone classes	Mahalanobis distance between perceptual cepstral parameters
(Founda et al. 2001)	Greek phrases with a single diphone replaced each time	Kullback-Leibler measure on power-normalised (LPC) spectra
(Tsuzaki & Kawai 2002)	438 Japanese cross-spliced stimuli	Measures using features extracted from computational model of the human auditory periphery

Table 3.1: Previous perceptual studies

3.2 Perceptual listening tests

As we reported in chapter 2, a perfect join cost should correlate highly with human perception of discontinuity at speech unit concatenation points. One way of designing an optimal join cost function is to use all possible distance measures on all varieties of speech parameterisation, then conduct a simple pair-wise comparison listening test to evaluate their performance and select the best join cost function for use in unit selection. However this design and evaluation is an iterative process and it is not practical to conduct listening tests for each and every possible join cost function.

A more efficient way would be to conduct a single perceptual listening test covering a wide range of synthetic stimuli. From this, we could obtain listeners' ratings and design the join cost function to maximise correlations with those ratings. Many of the previous studies presented in table 3.1 have used a similar methodology. However, these studies used mono syllabic isolated words. Here, we propose a variation on the usual perceptual test paradigm by deliberately including a **wide range of qualities** of join in **polysyllabic words**.

This test measures the degree of perceived concatenation discontinuity in whole (continuous speech) sentences generated by the state-of-the-art speech synthesis system, *rVoice*[1]. Brief details of this system and the voice used to synthesise these stimuli are given in Appendix A.

3.2.1 Test stimuli

A preliminary assessment indicated that spectral discontinuities are particularly prominent for joins in the middle of diphthongs, presumably because this is a point of spectral change (due to moving formant values). This study therefore focuses on such joins. Previous studies have also shown that diphthongs have higher discontinuity detection rates than long or short vowels (Syrdal 2001). Hence we feel, a join cost that is good for diphthongs should be good for all speech sounds (at least, all vowels).

We selected two natural sentences for each of five American English diphthongs: ey, ow, ay, aw and oy (in ARPAbet notation). One word in each sentence contained the

[1] An experimental version of rVoice was used in our research.

diphthong in a stressed syllable. The sentences for each diphthong (its corresponding IPA symbol in parentheses) are listed in table 3.2.

Diphthong	*Sentences*
ey (ei)	More **plac**es are in the pipeline. The government sought authori**zat**ion of his citizenship.
ow (oʊ)	European shares resist **glob**al fallout. The speech sym**pos**ium might begin on Monday.
ay (aɪ)	This is **high**ly significant. Primitive **trib**es have an upbeat attitude.
aw (aʊ)	A large **hous**ehold needs lots of appliances. Every picture is worth a **thous**and words.
oy(ɔi)	The **boy** went to play tennis. Never exp**loit** the lives of the needy.

Table 3.2: The stimuli used in the experiment. The syllable in bold contains the diphthong join. For each diphthong the corresponding IPA symbol is written in parentheses.

3.2.2 Test design

The sentences shown in table 3.2 were then synthesised using an experimental version of the *rVoice* speech synthesis system using an adult North-American male voice. The procedure for obtaining a number of synthetic versions for each sentence, to use them as test stimuli, is described in figure 3.1.

For each sentence, we made various synthetic versions by varying the two diphone candidates which make the diphthong, i.e. join B in figure 3.1, and keeping all the other units the same. Thus, we obtained around 600 different synthetic versions since there were around 25 candidates available for each diphone. We removed the synthetic versions which had a prominent join in either the phone immediately before or the phone immediately after the diphthong - marked as joins A and C in figure 3.1. The remaining versions were further pruned based on target features (pitch and duration) of the diphones making the diphthong, to ensure similar prosody among all synthetic

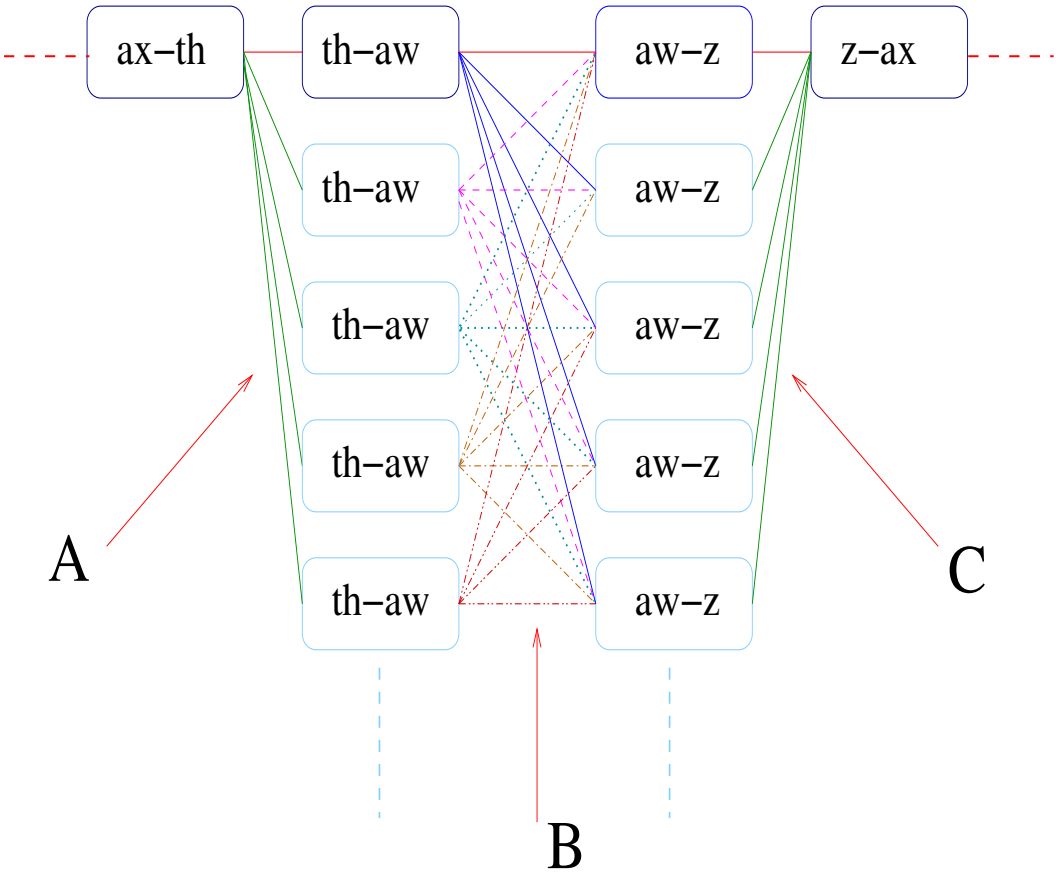

Figure 3.1: Perceptual listening test stimuli design

versions. This process resulted in around 30 versions with a wide range of concatenation discontinuities at the diphthong join (B in figure 3.1) and all with good joins elsewhere. We switched off smoothing to ensure no effect of it on concatenation discontinuities, though the quality of synthesiser is very high even without smoothing. This process was repeated for each sentence in table 3.2.

We manually selected what we judged to be the best and worst out of all these (around 30) versions. These were used to set the end points of the listener's rating scale, described in the following section.

3.2.3 Test procedure

There were 24 participants in our perceptual listening test; most of them were PhD or MSc students with some experience of speech synthesis. Most were native speakers of British English.

The test was carried out in blocks of around 35 test stimuli, with one block for each sentence in table 3.2. Subjects could take as long as they pleased over each block, and take rests between blocks. We also gave a choice to listeners either to complete all the test blocks in same session or over a few sessions, because of the difficulty involved in the task. Each test block contained a few duplications of some test stimuli to validate the listeners' ratings, discussed in section 3.2.4.1.

At the start of each test block, listeners were first shown the written sentence with an indication of which word contains the join. Then, they were presented with a pair of reference stimuli: one containing the best and the other the worst joins in order to set the end-points of a 1-to-5 scale. Listeners could listen to the reference stimuli as many times as they liked and they could also review them at regular intervals (for every 10 test stimuli) throughout the test.

They were then played each test stimulus in turn and were asked to rate the quality of that join on a scale of 1 (worst) to 5 (best). The five choices (1-5) supplied to listeners with a gloss by each number are as follows:

- 1 = *highly perceived join*
- 2 = *perceived join*

- 3 = *slightly perceived join*
- 4 = *not perceived, but not smooth*
- 5 = *no join at all*

They could listen to each test stimulus up to three times. Each test stimulus consisted of first the entire sentence, then only the word containing the join (extracted from the full sentence, not synthesised as an isolated word). On average, listeners took around an hour for completion of all test blocks. All of them took the tests on workstations in their quiet offices/laboratories and used headphones.

3.2.4 Test results

In table 3.3, we present the number of subjects for each sentence and the number of subjects with more than 50% consistency in rating the joins. There are different number of listeners for each sentence is because some listeners did not complete all the test blocks. The consistency check of all subjects was measured on a validation set, explained below.

Diphthong	*Number of subjects*	*Consistent subjects*
ey	20, 20	14, 11
ow	19, 17	10, 9
ay	24, 17	12, 9
aw	17, 19	17, 11
oy	19, 20	10, 11

Table 3.3: Consistency of subjects in listening tests, each number in a pair corresponds to the two sentences for each diphthong listed in table 3.2.

3.2.4.1 Consistency check

A validation set was included in the stimuli of each test block played to every listener to measure their consistency in rating them. This set consisted of a few duplications of some of the test stimuli. The size of the validation set was 6 out of 30 stimuli, thus

making each test block of 36 stimuli. The listener scores were cross-checked using the validation set.

If the preference ratings of a given sentence in the original set and the same sentence in the validation set are same, then we give a score of one, otherwise the score was considered as zero. Also, we introduced an intermediate score of 0.5 to avoid "just-miss" matches, i.e. if the subjects preferences differ on two identical stimuli by only 1 point on the 1-5 scale, we issued a score of 0.5. For example, if for a particular test stimuli, the listener rating is 3 on the one in the *original set* and 2 on the one in the *validation set*, then we gave a score of 0.5. Finally, for each listener, all ratings for each test stimuli in validation set are averaged and converted into a percentage. Listeners with a consistency of less than 50% are discarded for each test block in the computation of mean listener ratings.

The above consistency check will not catch listeners who always tended to give the same or very similar ratings to all stimuli (e.g. everything is rated 3 out of 5). Such listeners were found by manual inspection of the data, and removed.

The mean listener ratings used for correlating spectral distance measures (explained in the following sections) are only computed for the remaining listeners. The wide variation in number of consistent subjects per sentence may be due to the difficulty of the task; many listeners also commented that this was the case. For one sentence (the first sentence in the *aw* row) all subjects were consistent, but we feel that this is mere coincidence. Also we did not observe any relation between consistency and correlations between listener scores and our distance measures.

3.3 Correlations with statistical distances

The correlations were obtained between mean listener scores for each test sentence and distance measures computed by using various spectral features. As shown in the left-hand side plot of figure 3.2, if the mean listener scores and spectral distance measures have a strong linear relationship, i.e. if we can fit a straight line with a positive slope through these points with low mean squared error, then we can say that spectral distance measure and the perceptual scores are *highly correlated*.

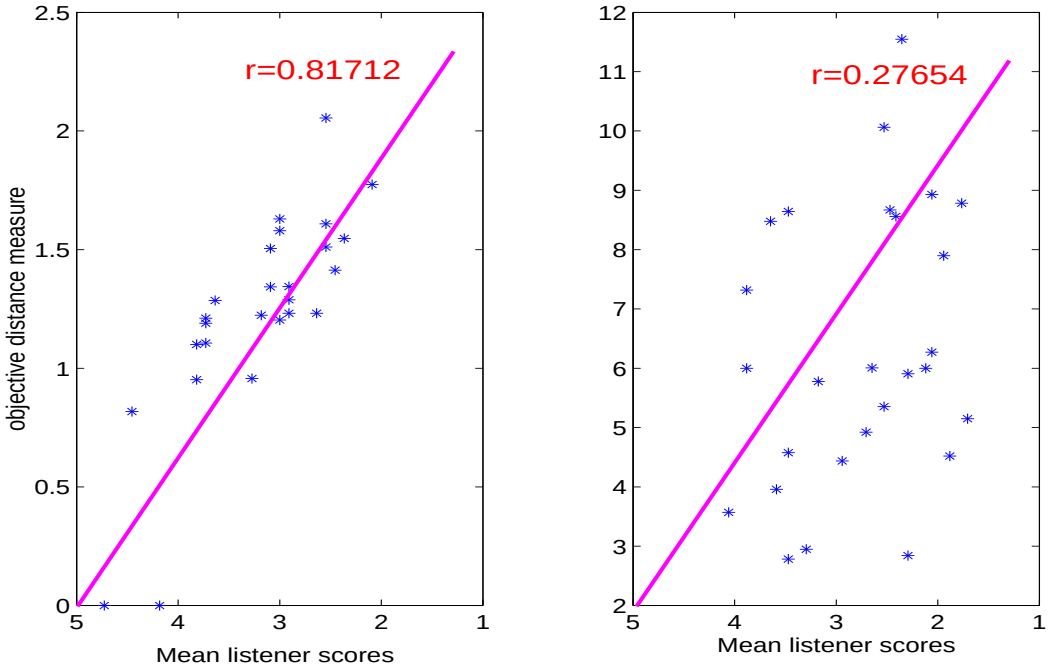

Figure 3.2: Example correlations

In contrast, as in the right-hand plot of figure 3.2, if points are sparsely placed from the straight line fit, then they have *low correlation*. We have computed the correlation coefficient, **r** using the following equation (Spiegel 1975).

$$\mathbf{r} = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \tag{3.1}$$

where n is number of pairs of data points, x and y. The value of **r** is always between -1.0 and +1.0. If the value of **r** is around 1.0, then there is a strong positive relation between two measures. Conversely, if the value of the correlation coefficient is around 0, then two measures are not correlated. A strong negative relationship is indicated by negative values of **r** around -1.0 . A good join cost function should have r values as close to 1 as possible.

In our case, since low join costs (in other words, spectral distance measures) should have high perceptual scores, the correlation coefficient should be close to -1. In figure 3.2, we have plotted the x-axis in reverse scale, i.e. 5-1 instead of 1-5, hence it shows positive slope. This is just for the reader's convenience. Throughout this thesis we compute correlation coefficients (r) as using equation 3.1 and present the $-r$ values

(e.g. for $r = -0.76$ it is taken as $r = 0.76$) for better readability.

Once the correlations are computed, a significance test can be used to determine the probability that the observed correlation occurred is a real one and not by chance. A one-tailed test was chosen, since we know the direction of the relationship between spectral distance measures and perceptual scores: a low perceptual score corresponds to a high spectral distance. We computed the t statistic using (Spiegel 1975):

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{3.2}$$

where n is the length of each test of block. Once we have the t statistic, we can refer to Student's t distribution table to find the significance value (α) of the test.

Actually, we are more interested in knowing the significance of correlations. So, we observed out of our ten cases, how many times various spectral distance measures produced significant correlations. Then, we compared the distance measures that had significant correlations with the listener ratings. This is because, one of our goals is to generalise the distance measure for all the phones. Though we tested distance measures only on diphthong joins, our hypothesis is that if a distance measure or join cost function works for diphthongs, which have difficult joins, then it will perform well for all other phones.

Furthermore, the join cost functions that perform well on a *large* number of cases of diphthongs are expected to generalise better to other phone classes. We assumed that our ten cases are independent and obtained the distance measure for each case independently: a successful distance measure has significant correlation with mean listener ratings. We also compare various distance measures using the global correlation coefficient, which is calculated using perceptual scores of all our stimuli and corresponding distances.

3.3.1 Correlations for single frame distances

Correlation coefficients of various spectral distance measures with mean listener preference ratings (perceptual scores) are reported in tables 3.4, 3.5, 3.6 and 3.7. The first row and the second row correspond to the first sentence and the second sentence for each diphthong in table 3.2. The correlation coefficients above the 1% significant level are in bold face. To show what correlation plots look like, we present those for the MFCC distance measures and mean listener scores in the case of the second sentence of the diphthong *aw* in figure 3.3.

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	0.27	0.29	0.27	0.29	0.19	0.29
	0.65	**0.64**	**0.69**	**0.67**	**0.73**	**0.65**
ow	0.31	0.28	0.31	0.24	0.31	0.18
	0.57	**0.57**	**0.55**	**0.54**	**0.59**	**0.59**
ay	0.32	0.31	0.34	0.29	**0.43**	0.28
	0.61	**0.63**	**0.62**	**0.66**	**0.56**	**0.57**
aw	0.33	0.31	0.35	0.29	0.26	0.16
	0.76	**0.76**	**0.73**	**0.74**	**0.77**	**0.78**
oy	0.08	0.08	0.08	0.12	0.25	0.19
	0.14	0.17	0.13	0.19	0.16	0.25

Table 3.4: Correlation between perceptual scores and various distance measures computed using MFCCs

From tables 3.4, 3.5, 3.6 and 3.7, it is clear that no distance measure performs well in all cases. The distance measures using LSFs have a higher number of significant correlations compared to those using MFCCs or MCAs or LPC spectra. Unfortunately, none of these measures yield 1% significant level correlations for two of our ten sentences (first sentences in *ey* and *ow* rows). Including the delta coefficients (explained in section 2.3.4) improved correlations in some cases, but not all. One has to consider the increase

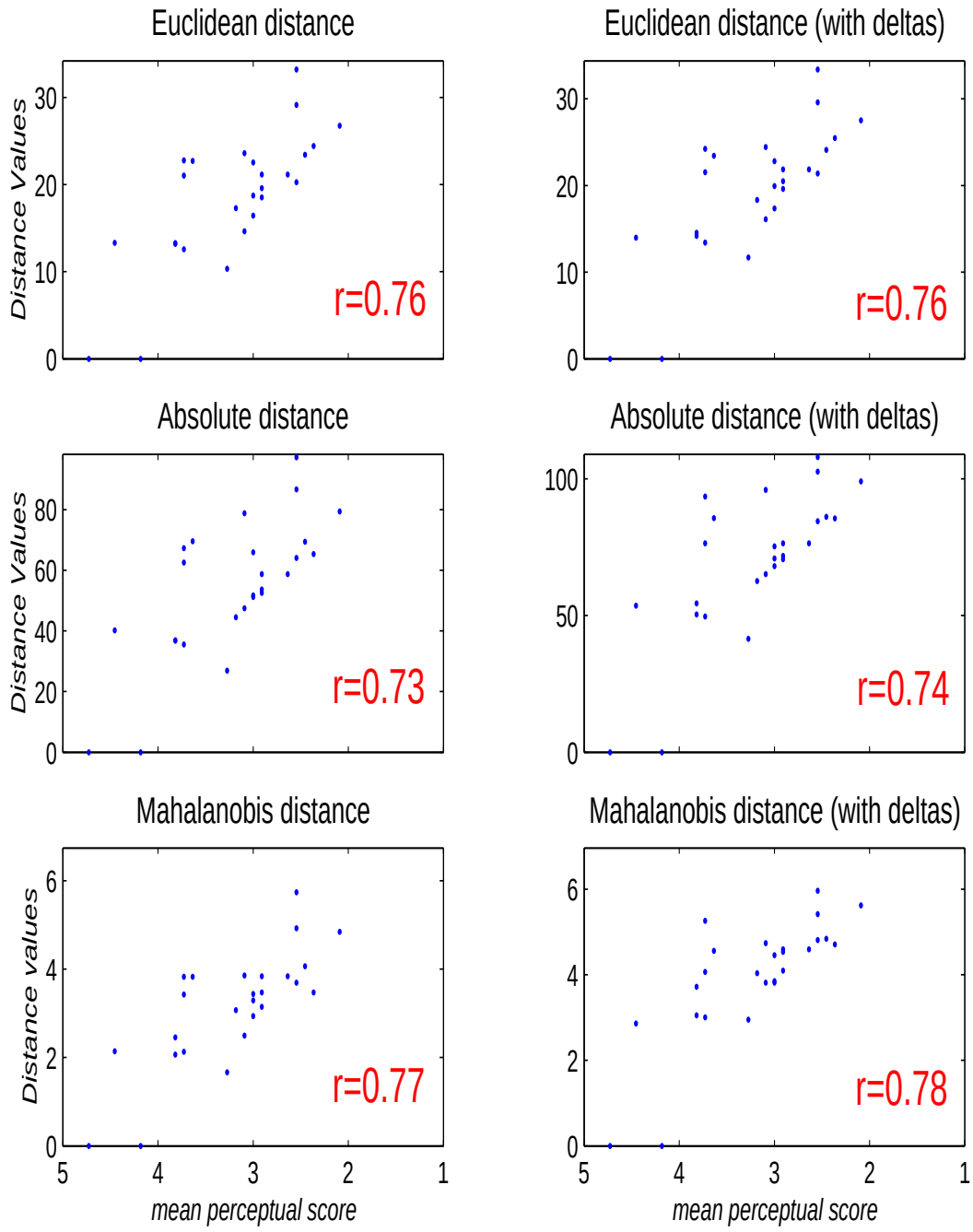

Figure 3.3: Correlation plots of various distances using MFCCs for the second case of *aw* join

in memory requirements, i.e. double the original size, when using delta coefficients in distance computation. Spectral distance measures using the Mahalanobis distance have high correlations for more cases, followed by symmetric K-L distance. Also, a simple absolute distance measure performs as well as any distance measure in some cases.

In table 3.4, we present correlations with perceptual scores and Euclidean, absolute and Mahalanobis distance measures computed using MFCCs and delta MFCCs. In many cases all these distances have four 1% significant correlations. The **Mahalanobis distance using MFCCs** alone has more (**five** out of ten cases) 1% significant correlations than other distance measures using MFCCs.

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*		*K-L*
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.04	0.12	0.14	0.24	0.14	0.27	0.30
	0.73	**0.70**	**0.75**	**0.72**	**0.72**	**0.58**	**0.73**
ow	0.40	0.24	0.35	0.16	0.39	0.12	0.33
	0.53	**0.57**	**0.46**	**0.48**	0.42	**0.46**	**0.45**
ay	0.16	0.07	0.12	0.02	0.17	0.01	0.38
	0.50	**0.61**	**0.51**	**0.65**	**0.47**	**0.46**	**0.63**
aw	0.21	0.35	0.09	0.37	0.23	**0.57**	0.16
	0.80	**0.80**	**0.79**	**0.78**	**0.78**	**0.65**	**0.79**
oy	0.18	0.22	0.11	0.14	0.28	0.27	0.15
	0.12	0.29	0.17	0.30	0.23	**0.48**	0.30

Table 3.5: Correlation between perceptual scores and various distance measures computed using LSFs

Distance measures using LSFs have higher (i.e. stronger) correlations in more cases, as observed in table 3.5 than any of our speech parameterisations. In particular, **Mahalanobis distance computed on LSFs and their delta coefficients** has the highest number (**six** out of ten cases) of 1% significant correlations of any single frame distance presented in this chapter.

From table 3.6, it is clear that distance measures using MCA coefficients correlate

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*		*K-L*
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.36	0.29	0.34	0.34	0.35	0.35	**0.45**
	0.62	**0.49**	**0.62**	**0.50**	0.41	**0.49**	**0.65**
ow	0.01	0.11	0.06	0.12	0.14	-0.00	0.11
	0.39	**0.49**	0.40	**0.52**	**0.52**	**0.56**	0.32
ay	-0.02	0.06	-0.03	0.01	0.01	0.04	0.11
	0.41	0.22	0.38	0.24	0.40	**0.48**	0.45
aw	**0.46**	0.28	0.36	0.36	**0.47**	**0.54**	0.36
	0.76	**0.54**	**0.75**	**0.54**	**0.76**	**0.77**	**0.82**
oy	0.28	**0.51**	0.24	**0.54**	0.21	0.37	0.16
	0.12	0.24	0.14	0.36	0.27	0.39	0.27

Table 3.6: Correlation between perceptual scores and various distance measures computed using MCAs

well with perceptual scores in only a few cases compared to those using MFCCs and LSFs. This is contrary to our previous results (Vepa et al. 2002*b*), where we observed many 1% significant correlations. The results presented here are based on perceptual data from more listeners, and are therefore more reliable.

However, MCAs do have an advantage over MFCCs, LSFs and LPC spectra due to their compact size: the size[2] of the MCA feature vector is only 12, whereas MFCCs are 26, LSFs are 24 and LPCs are 13 (but, involves computation of the discrete spectrum). Thus, using MCA coefficients reduces the memory required quite significantly. Considering this, the **Mahalanobis distance measure using MCA coefficients with their deltas** is also a good choice, which has **five** 1% significant correlations out of ten cases. This is also holds for MFCCs using the Mahalanobis distance, which in turn produces the same (five) results.

We also computed distances based on LPC spectra, as some researchers found in their studies that symmetrical Kullback-Leibler (K-L) distance on LPC spectra has the

[2]all these figures include delta coefficients

Diphthong	*K-L*	*I-S*	*COSH*
ey	0.35	-0.04	-0.15
	0.68	-0.20	-0.42
ow	0.38	-0.19	0.04
	0.44	0.08	0.12
ay	0.34	-0.20	0.08
	0.60	0.41	0.36
aw	-0.40	0.09	-0.14
	0.70	-0.22	0.09
oy	-0.14	0.32	0.32
	0.30	0.07	-0.21

Table 3.7: Correlation between perceptual scores and various distance measures based on LPC power spectra.

highest prediction rates (Klabbers & Veldhuis 1998, Founda et al. 2001). table 3.7 shows correlations between perceptual scores and K-L, Itakura-Saito (I-S) and COSH distances computed using LPC spectra. Only the K-L distance on LPC spectra has a few (3 out of 10) significant correlations, other distances do not produce a single significant correlation. They have the worst performance of any of our spectral distance measures. Also, it is worth noting that the computation and memory requirements of these distances is very high as we also need to calculate spectra. Hence we are not including these distances in our further analysis of spectral distance measures.

We also computed the global correlations for various distance measures obtained using MFCCs, LSFs and MCA coefficients and presented in table 3.8. In table 3.8 we also present the number of 1% significant correlations for each case. The global correlation values are high for K-L and Mahalanobis distances computed using LSFs and also for Mahalanobis distance computed using MFCCs.

Distance measure	MFCC		LSF		MCA	
	1% sig.	Global corr.	1% sig.	Global corr.	1% sig.	Global corr.
Euclidean	4	0.25	4	0.26	3	0.15
Euclidean (deltas)	4	0.25	4	0.27	4	0.24
Absolute	4	0.27	4	0.26	2	0.17
Absolute (deltas)	4	0.24	4	0.26	4	0.26
Mahalanobis	5	0.33	3	0.35	3	0.27
Mahalanobis (deltas)	4	0.27	6	0.28	5	0.25
Kullback-Leibler	N/A	N/A	4	0.38	3	0.25

Table 3.8: Number of 1% significant correlations and global correlation coefficients for single-frame spectral distances

3.3.2 Correlations for multi-frame distances

We computed spectral distances using multiple frames of the two units on either side of the concatenation. For example, to compute a three frame spectral distance, we use the middle frame at the join and one frame on either side of the units at the join. This method is described in figure 3.4.

Consider A and C are neighbouring frames of join frame B in the first unit and similarly, B' is the join frame for the second unit with A' and C' as adjacent frames. In order to compute the spectral distance between these two units using three frames, first we calculate individual frame distances:

1. Distance between frame A and frame A' (shown as D1 in figure 3.4)

2. Distance between frame B and frame B' (D2 in figure 3.4, this is same as our single frame distance described in section 2.2.1)

3. Distance between frame C and frame C' (D3 in figure 3.4)

Then we average the three distances, ($D1, D2$ and $D3$) respectively, and use this as the spectral distance measure between two units. Hence, these distances are also termed as **average multi-frame distances**.

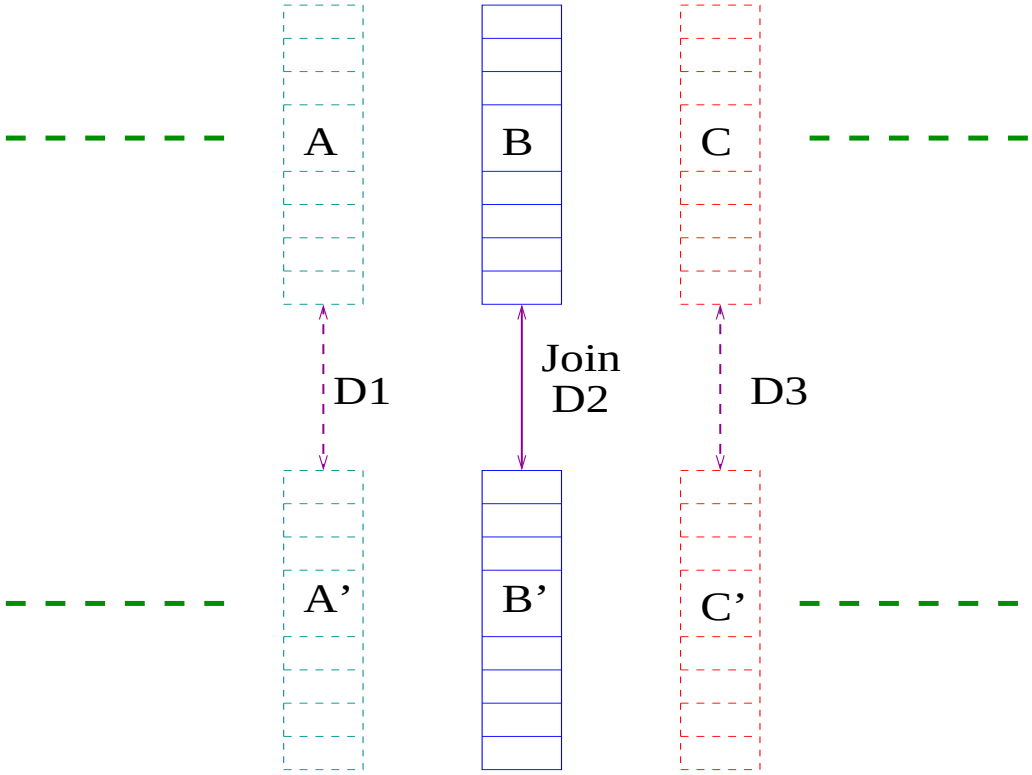

Figure 3.4: Multi (three) frame distances

3.3.2.1 Three-frame spectral distances

Tables 3.9, 3.10 and 3.11 show correlations with perceptual scores and three-frame distances computed using MFCCs, LSFs and MCA coefficients. The correlation results for these distances computed for MFCC are similar to those of single frame distances except in one column (Mahalanobis distance using MFCCs), where there are more sig-

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	0.29	0.32	0.29	0.34	0.20	0.32
	0.72	**0.71**	**0.73**	**0.71**	**0.78**	**0.69**
ow	0.26	0.25	0.26	0.21	0.26	0.18
	0.57	**0.57**	**0.54**	**0.54**	**0.57**	**0.55**
ay	0.32	0.30	0.32	0.25	0.41	0.23
	0.68	**0.71**	**0.67**	**0.75**	**0.63**	**0.74**
aw	0.31	0.33	0.31	0.33	0.21	0.30
	0.79	**0.80**	**0.79**	**0.80**	**0.79**	**0.80**
oy	0.05	0.07	0.06	0.10	0.21	0.24
	0.17	0.20	0.15	0.20	0.20	0.29

Table 3.9: Correlation between perceptual scores and various distance measures computed using **three** frames of **MFCC**s

nificant correlations for single frame distances. In case of LSFs, correlation results with multi-frame distances are not better than those of single frame distances.

From the comparison of correlations obtained from three-frame distances and single frame distances of MCA coefficients, we can observe more significant correlations for three-frame distances. We have achieved **six** 1% significant correlations for **Mahalanobis distance using MCA coefficients plus deltas** out of our ten cases. But, one has to pay the price in terms of substantial disk usage, i.e. a three fold increase in memory requirements compared to single frame distances, as we need to store six frames (three at the start and three at the end) for each unit to compute spectral distance

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*		*K-L*
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.04	0.14	0.12	0.27	0.13	0.37	0.29
	0.75	**0.74**	**0.77**	**0.75**	**0.75**	**0.63**	**0.74**
ow	0.35	0.24	0.33	0.21	0.37	0.17	0.29
	0.53	**0.47**	**0.47**	0.44	0.44	0.29	**0.48**
ay	0.13	0.05	0.08	-0.00	0.12	0.01	0.37
	0.51	**0.68**	**0.52**	**0.72**	**0.48**	**0.57**	**0.66**
aw	0.23	0.38	0.14	0.34	0.27	**0.57**	0.17
	0.80	**0.78**	**0.79**	**0.77**	**0.78**	**0.67**	**0.80**
oy	0.22	0.24	0.17	0.22	0.34	0.41	0.17
	0.15	0.28	0.18	0.28	0.22	0.42	0.33

Table 3.10: Correlation between perceptual scores and various distance measures computed using **three** frames of **LSF**s

measures.

3.3.2.2 Five-frame spectral distances

We also computed five-frame spectral distances, i.e. using two frames of spectral parameters on either side of the join in addition to the frame at the join. Correlation coefficients are presented in tables 3.12, 3.13 and 3.14 for MFCC, LSF and MCA parameters respectively. Again, we have not observed any improvement in the case of MFCCs and LSFs compared to the single frame distances, in fact these results are worse. But, for MCAs we achieved **six** 1% significant correlations for **Mahalanobis distance using MCA coefficients plus deltas**.

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*		*K-L*
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.40	0.23	0.39	0.30	0.32	0.34	**0.48**
	0.65	**0.47**	**0.67**	**0.50**	0.44	**0.47**	**0.65**
ow	0.31	0.22	0.33	0.28	0.32	0.29	0.25
	0.48	**0.47**	**0.46**	**0.53**	**0.55**	**0.57**	0.41
ay	-0.03	0.08	-0.04	0.03	0.04	0.05	0.12
	0.48	0.22	0.44	0.25	**0.47**	**0.55**	**0.55**
aw	**0.46**	0.22	0.38	0.30	**0.57**	**0.62**	0.40
	0.76	**0.51**	**0.78**	**0.52**	**0.72**	**0.72**	**0.80**
oy	0.34	**0.49**	0.33	**0.54**	0.35	0.31	0.23
	0.16	0.35	0.19	**0.50**	0.34	**0.78**	0.32

Table 3.11: Correlation between perceptual scores and various distance measures computed using **three** frames of **MCA**s

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	0.42	0.43	0.40	0.44	0.32	0.41
	0.72	**0.72**	**0.72**	**0.71**	**0.77**	**0.69**
ow	0.27	0.26	0.25	0.23	0.26	0.21
	0.57	**0.58**	**0.54**	**0.55**	**0.57**	**0.58**
ay	0.33	0.32	0.33	0.28	**0.42**	0.28
	0.73	**0.75**	**0.72**	**0.77**	**0.66**	**0.73**
aw	0.27	0.30	0.26	0.32	0.17	0.34
	0.81	**0.81**	**0.80**	**0.80**	**0.81**	**0.78**
oy	0.01	0.05	0.03	0.10	0.14	0.25
	0.20	0.22	0.17	0.22	0.23	0.31

Table 3.12: Correlation between perceptual scores and various distance measures computed using **five** frames of **MFCC**s

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*		*K-L*
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.05	0.17	0.13	0.27	0.14	0.36	0.31
	0.74	**0.71**	**0.77**	**0.72**	**0.75**	**0.62**	**0.73**
ow	0.32	0.23	0.29	0.20	0.34	0.18	0.27
	0.52	**0.49**	**0.45**	**0.46**	**0.47**	0.39	**0.46**
ay	0.11	0.07	0.07	0.02	0.09	0.04	0.37
	0.59	**0.73**	**0.58**	**0.77**	**0.55**	**0.66**	**0.71**
aw	0.25	0.41	0.16	0.39	0.31	**0.59**	0.15
	0.81	**0.80**	**0.80**	**0.79**	**0.79**	**0.71**	**0.80**
oy	0.25	0.31	0.23	0.35	0.35	0.45	0.21
	0.16	0.24	0.18	0.24	0.21	0.33	0.34

Table 3.13: Correlation between perceptual scores and various distance measures computed using **five** frames of **LSF**s

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*		*K-L*
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.38	0.23	0.39	0.30	0.28	0.33	**0.46**
	0.69	**0.49**	**0.73**	**0.52**	**0.51**	**0.50**	**0.67**
ow	0.33	0.20	0.35	0.27	0.32	0.29	0.28
	0.52	**0.48**	**0.48**	**0.53**	**0.57**	**0.59**	0.43
ay	-0.04	-0.02	-0.05	-0.05	0.02	0.03	0.13
	0.54	0.24	**0.49**	0.26	**0.51**	**0.55**	**0.60**
aw	0.40	0.19	0.31	0.24	**0.52**	**0.56**	0.30
	0.76	**0.52**	**0.79**	**0.54**	**0.72**	**0.71**	**0.79**
oy	0.43	**0.50**	0.40	**0.54**	0.45	0.44	0.31
	0.18	0.41	0.19	**0.51**	0.35	**0.77**	0.34

Table 3.14: Correlation between perceptual scores and various distance measures computed using **five** frames of **MCA**s

3.3.2.3 Seven-frame spectral distances

Similarly, we have carried out the computation of seven-frame distances. Tables 3.15, 3.16 and 3.17 show the correlations with mean listener scores and seven-frame distances computed using MFCCs, LSFs and MCA coefficients. We have observed more 1% significant correlations in the case of MCA parameters, **seven** for Mahalanobis distance and **six** for Mahalanobis distance computed with delta coefficients also.

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	0.40	0.41	0.38	0.42	0.30	0.39
	0.73	**0.72**	**0.73**	**0.71**	**0.76**	**0.69**
ow	0.26	0.24	0.24	0.22	0.27	0.19
	0.59	**0.60**	**0.58**	**0.59**	**0.60**	**0.61**
ay	0.34	0.32	0.33	0.27	0.41	0.27
	0.74	**0.76**	**0.73**	**0.77**	**0.67**	**0.74**
aw	0.23	0.26	0.24	0.28	0.15	0.29
	0.82	**0.82**	**0.82**	**0.81**	**0.82**	**0.78**
oy	0.05	0.12	0.05	0.17	0.14	0.34
	0.21	0.23	0.19	0.22	0.23	0.28

Table 3.15: Correlation between perceptual scores and various distance measures computed using **seven** frames of **MFCC**s

We present the number of 1% significant correlations and global correlation values for seven-frame distances in table 3.18. The global correlations are high for K-L distance computed using LSFs followed by Mahalanobis distance using MFCCs and Mahalanobis distance using LSFs.

3.3.2.4 Nine-frame spectral distances

Finally, we computed nine-frame distances, i.e. using four frames on either side of the join and the frame at the join. The correlation results for MFCC, LSF and MCAs for

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*		*K-L*
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.06	0.21	0.14	0.30	0.16	0.39	0.33
	0.74	**0.71**	**0.76**	**0.72**	**0.75**	**0.63**	**0.73**
ow	0.29	0.20	0.27	0.20	0.33	0.17	0.25
	0.52	**0.48**	0.45	**0.46**	**0.49**	0.39	**0.45**
ay	0.08	0.03	0.05	0.00	0.05	-0.01	0.36
	0.60	**0.75**	**0.60**	**0.77**	**0.57**	**0.68**	**0.71**
aw	0.26	0.44	0.17	0.41	0.32	**0.61**	0.15
	0.82	**0.81**	**0.82**	**0.81**	**0.81**	**0.73**	**0.81**
oy	0.26	0.33	0.25	0.39	0.38	0.47	0.22
	0.16	0.23	0.18	0.23	0.19	0.31	0.34

Table 3.16: Correlation between perceptual scores and various distance measures computed using **seven** frames of **LSF**s

these distances are demonstrated in tables 3.19, 3.20 and 3.21. Here also we obtained **seven** 1% significant correlations **Mahalanobis distance using MCA coefficients**. But other correlations are not as good as those of single frame distances. As we said earlier, one has to consider the amount of disk usage and computation requirements before employing these distance measures for join cost computation.

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*		*K-L*
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.36	0.22	0.39	0.29	0.23	0.35	**0.46**
	0.68	**0.50**	**0.72**	**0.53**	**0.55**	**0.54**	**0.65**
ow	0.36	0.21	0.38	0.25	0.34	0.28	0.31
	0.56	**0.56**	**0.52**	**0.60**	**0.59**	**0.58**	**0.48**
ay	-0.04	-0.04	-0.05	-0.06	-0.00	0.01	0.13
	0.57	0.24	**0.54**	0.27	**0.54**	**0.59**	**0.64**
aw	0.38	0.21	0.32	0.25	**0.49**	**0.49**	0.27
	0.73	**0.53**	**0.75**	**0.54**	**0.70**	**0.69**	**0.77**
oy	0.41	0.45	0.40	**0.50**	**0.51**	0.26	0.37
	0.41	**0.53**	0.42	**0.61**	**0.51**	**0.74**	0.43

Table 3.17: Correlation between perceptual scores and various distance measures computed using **seven** frames of **MCA**s

Distance	MFCC		LSF		MCA	
measure	1% sig.	Global corr.	1% sig.	Global corr.	1% sig.	Global corr.
Euclidean	4	0.30	4	0.27	4	0.20
Euclidean (deltas)	4	0.30	4	0.28	4	0.22
Absolute	4	0.30	3	0.27	4	0.22
Absolute (deltas)	4	0.28	4	0.27	5	0.25
Mahalanobis	4	0.37	4	0.35	7	0.33
Mahalanobis (deltas)	4	0.34	4	0.31	6	0.31
Kullback-Leibler	N/A	N/A	4	0.38	5	0.30

Table 3.18: Number of 1% significant correlations and global correlation coefficients for seven-frame spectral distances

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	0.40	0.41	0.37	0.42	0.32	0.40
	0.75	**0.73**	**0.74**	**0.71**	**0.77**	**0.68**
ow	0.26	0.25	0.25	0.24	0.28	0.23
	0.58	**0.58**	**0.57**	**0.56**	**0.59**	**0.56**
ay	0.35	0.33	0.34	0.28	0.41	0.28
	0.75	**0.77**	**0.73**	**0.77**	**0.67**	**0.74**
aw	0.21	0.22	0.21	0.22	0.15	0.20
	0.83	**0.82**	**0.83**	**0.82**	**0.83**	**0.79**
oy	0.17	0.22	0.17	0.25	0.23	0.36
	0.22	0.23	0.21	0.23	0.23	0.28

Table 3.19: Correlation between perceptual scores and various distance measures computed using **nine** frames of **MFCC**s

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*		*K-L*
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.09	0.23	0.16	0.31	0.20	0.39	0.36
	0.74	**0.67**	**0.76**	**0.69**	**0.75**	**0.56**	**0.72**
ow	0.33	0.24	0.31	0.24	0.36	0.21	0.27
	0.49	**0.45**	0.42	0.42	**0.47**	0.32	0.43
ay	0.08	0.04	0.07	0.02	0.04	-0.01	0.36
	0.62	**0.74**	**0.62**	**0.76**	**0.60**	**0.70**	**0.71**
aw	0.28	0.36	0.19	0.32	0.33	**0.47**	0.16
	0.84	**0.82**	**0.84**	**0.82**	**0.83**	**0.74**	**0.81**
oy	0.28	0.26	0.25	0.32	0.42	0.26	0.24
	0.14	0.23	0.16	0.23	0.15	0.32	0.33

Table 3.20: Correlation between perceptual scores and various distance measures computed using **nine** frames of **LSF**s

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*		*K-L*
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.18	0.17	0.23	0.23	0.16	0.24	0.39
	0.67	**0.52**	**0.71**	**0.53**	**0.56**	**0.52**	**0.63**
ow	0.43	0.30	0.44	0.32	0.39	0.37	0.33
	0.54	**0.48**	**0.50**	**0.49**	**0.54**	0.42	**0.47**
ay	-0.03	0.02	-0.04	-0.02	-0.02	0.01	0.15
	0.61	0.25	**0.58**	0.28	**0.59**	**0.64**	**0.67**
aw	0.36	0.20	0.31	0.23	**0.48**	0.39	0.25
	0.74	**0.54**	**0.76**	**0.55**	**0.72**	**0.72**	**0.77**
oy	0.44	0.45	0.44	**0.52**	**0.56**	0.31	0.36
	0.51	**0.56**	**0.50**	**0.63**	**0.54**	**0.73**	**0.48**

Table 3.21: Correlation between perceptual scores and various distance measures computed using **nine** frames of **MCA**s

3.4 Analysis of Variance (ANOVA)

In the previous section, we presented correlations obtained between mean listener scores and distance measures computed by using various spectral features, and also checked the significance of these correlation values. Though we can observe some strong and significant correlations, it is also worthwhile to look our experimental data from another statistical perspective to gain more insight.

We have chosen analysis of variance (ANOVA) as our second statistical procedure. ANOVA is a method by which the source of variability is identified. The key statistic in this method is the F-test of differences between means of groups of two or more formed by values of the factor[3]. If the group means do not differ significantly then it can be inferred that the factor did not have an effect on the measured value. For example, the factor can be a speech parameterisation and the measured value is a spectral distance

[3]A variable with one or more values/categories thought to be a cause of variation in a measured value

obtained using various spectral features..

We carried out ANOVA experiments on our perceptual data and single-frame spectral distances. In our case, the sources of variation or factors are:

1. The four distance measures; *Euclidean, absolute, Mahalanobis* and *Kullback-Leibler (K-L)*

2. The six speech parameterisations (or) spectral features; *MFCC, MFCC with deltas, LSF, LSF with deltas, MCA* and *MCA with deltas*

3. The five diphthongs; *ey, ow, ay, aw* and *oy*

4. The ten test sentences, listed in table 3.2

5. The variation between different listeners in their response to test stimuli

3.4.1 ANOVA on spectral distances

First, we a performed one-way ANOVA within-subjects design by grouping the spectral distances for each speech parameterisation (listed above) based on the distance measures used, i.e. Euclidean, absolute, Mahalanobis and K-L. The F and p-values are shown in table 3.22. The p-value is zero for all cases, which means that all four distance measures are significantly different. In other words, there is a strong indication that the spectral distances computed using these four distance measures are not the same.

Spectral features	*F*	*p-value*
MFCCs	934.93	0
MFCCs+deltas	1142.00	0
LSFs	356.21	0
LSFs+deltas	550.70	0
MCAs	102.21	0
MCAs+deltas	481.43	0

Table 3.22: One-way ANOVA on four distance measures computed on various spectral features

Table 3.23 presents the results of two-way ANOVA on spectral distances with two factors, namely distance measures and spectral features. Here also all the p-values are zero. Hence, there is a strong indication that the spectral distances vary from one distance measure to another and also from one spectral feature to another. Also, the effect of distance measure is much stronger than that due to spectral feature. There is also a significant interaction between distance measures and spectral features, i.e. the effect of distance measure on spectral distance varies according to the spectral feature used.

Source	*F*	*p-value*
Distance measures	4084.70	0
Spectral features	56.76	0
Distance measures * Spectral features	24.09	0

Table 3.23: Two-way ANOVA on spectral distances, factors are distance measures and spectral features.

The results from three-way ANOVA on spectral distances are shown in table 3.24. Here we used a full model, i.e. including interactions of all levels. However, all the sources in the table do not have full rank due to missing factor combinations. Actually, our data contains three distances (obtained using Euclidean, absolute and Mahalanobis distance measures) for MFCC and MFCC+delta features, where as other features (LSF, LSF+delta, MCA and MCA+delta features) used four distances including Kullback Leibler(K-L) distance measure. Hence, it is impossible to estimate the three-way interaction effects.

We removed interactions containing features as a factor, i.e. sentences*features and sentences*features*distances from the full model (presented before) and carried out three-way ANOVA once again. The results are presented in table 3.24. The p-values are zero for all main effects and interaction effects, which indicate that these three factors differ significantly, i.e. spectral distances relate to all three factors. There is an interaction between sentences and spectral features, also between sentences and distance measures.

Table 3.25 presents the results of three-way ANOVA on spectral distances with

Source	*F*	*p-value*
Sentences	11.99	0
Spectral features	794.08	0
Distance measures	823.72	0
Sentences * Spectral features	4.68	0
Sentences * Distance measures	3.83	0

Table 3.24: Three-way ANOVA on spectral distances without a full model, factors are sentences, spectral features and distance measures.

factors as diphthongs, spectral features and distance measures. This is the same as above except the first factor is changed to diphthongs from sentences. Here also we observed that the three factors differ significantly and there are interactions between diphthongs and spectral features and also between diphthongs and distance measures.

Source	*F*	*p-value*
Diphthongs	19.41	0
Spectral features	789.47	0
Distance measures	818.30	0
Diphthongs * Spectral features	7.57	0
Diphthongs * Distance measures	6.24	0

Table 3.25: Three-way ANOVA on spectral distances without a full model, factors are diphthongs, spectral features and distance measures.

Finally, we carried out four-way ANOVA on spectral distances with factors as diphthongs, sentences, spectral features and distance measures. The sentence factor used here is different in its meaning from other tests, this has only two categories, i.e. first sentence or second sentence of given diphthong. The F and p-values are presented in table 3.26. The p-value for sentences is not significant, which implies that the spectral distances would not be sensitive to whether the first or second sentence was presented. Surprisingly, this is contrary to the observations made from tables 3.4, 3.5,3.6 and 3.7, where we always observed significant correlations for the second sentences of each diph-

thong. The spectral distances are related to the other three factors significantly, i.e. spectral distances vary from one diphthong to another, one spectral feature to another and one distance measure to another.

Source	*F*	*p-value*
Diphthongs	21.61	0
Sentences	1.56	0.2119
Spectral features	758.13	0
Distance measures	785.27	0

Table 3.26: Four-way ANOVA on spectral distances without a full model,factors are diphthongs, sentences, spectral features and distance measures.

3.4.2 ANOVA on perceptual data

We performed one-way ANOVA on listeners' scores of each test block[4], the data is grouped for each listener. The number of groups for each test block varies as some listeners did not complete all the test blocks. The p-values for all blocks and corresponding F values are presented in table 3.27 and these values indicate that listeners' scores differ significantly.

Another one-way ANOVA on listeners' scores was performed with a different factor compared to previous design. Here the groups were formed for each test stimulus of each of the ten test blocks. The F and p-values for each test block are shown in table 3.28, which indicate that the test stimuli are significantly different with 1% significance level ($\alpha = 0.01$) in all test blocks except for the second sentence of *ow*, where the significance at 5% significance level ($\alpha = 0.05$).

We also carried out two-way ANOVA on listeners' scores, where the factors are sentences and listeners, the F and p-values are presented in table 3.29. Here, we are not able do sentences*listeners interaction due to missing factor combinations: in our data number of listeners varies for each sentence. The p-values indicate that listeners' scores differ from one sentence to another and also one listener to another.

[4]As mentioned in 3.2.3 each block has around 35 test stimuli and there are ten such blocks in our data, i.e. one block for each sentence in table 3.2

Diphthong	*F*	*p-value*
ey	4.11	0
	5.69	0
ow	4.31	0
	6.11	0
ay	5.33	0
	7.26	0
aw	6.41	0
	8.78	0
oy	13.49	0
	7.05	0

Table 3.27: One-way ANOVA on listeners' scores obtained from our perceptual tests, groups are listeners

Diphthong	*F*	*p-value*
ey	8.22	0
	4.18	0
ow	2.85	0
	1.54	0.05
ay	3.08	0
	3.13	0
aw	8.35	0
	4.01	0
oy	2.33	0.0009
	3.93	0

Table 3.28: One-way ANOVA on listeners' scores obtained from our perceptual tests, groups are test stimuli

In table 3.30, we present results obtained from two-way ANOVA on listeners' scores with factors as diphthongs and listeners. Here also, we observed that the listeners' scores differ from one diphthong to another and also one listener to another.

Source	*F*	*p-value*
Sentences	10.27	0
Listeners	5.73	0

Table 3.29: Two-way ANOVA on listeners' scores obtained from our perceptual tests, factors are sentences and listeners.

Source	*F*	*p-value*
Diphthongs	14.06	0
Listeners	6.52	0

Table 3.30: Two-way ANOVA on listeners' scores obtained from our perceptual tests, factors are diphthongs and listeners.

3.5 Conclusions

Finding a spectral distance measure to use in the join cost function, which correlates well with human listeners' perception of join discontinuity, is difficult. The correlations between perceptual data and various join costs based on four speech parameterisations – Mel frequency cepstral coefficients (MFCCs), line spectral frequencies (LSFs), multiple centroid analysis (MCA) coefficients and LPC spectra – and four distance measures – Euclidean, absolute, Mahalanobis and symmetric Kullback-Leibler (K-L) (also Itakura-Saito and COSH distances for LPC spectra) – were computed using frames at the join. Spectral distances were also computed using speech parameters (MFCC/LSF/MCA) of multi-frames (three/five/seven/nine) of the two units on either side of the join.

Although there is no significant difference among all these correlations, we suggest that a distance measure which performs well in a greater number of diphthong cases is likely also to perform better across a greater number of other phoneme classes, where the difference is more likely to be significant. Hence, we compared various distance measures based on the number of 1% significant correlations.

Among single frame distances, the Mahalanobis distance using LSFs plus delta coefficients has the highest number (six out of ten cases) of 1% significant correlations. If we consider the storage requirements of the inventory then Mahalanobis distance using

MCAs and their delta coefficients and similar distance measures computed on MFCCs alone are also reasonable choices as both of them have five 1% significant correlations. Also these distance measures have high global correlation coefficients compared to other single frame distance measures.

It is evident that single frame distances result in better correlations than those of multi-frame distances in most of the cases, also they are compact in size. Only distances computed using MCAs have more 1% significant correlations, **six** for Mahalanobis distance using delta coefficients in all multi-frame distances (except nine-frame distance). We also achieved **seven** 1% significant correlations for Mahalanobis distance using MCAs for seven and nine frame distances. The better performance of MCAs than LSFs and MFCCs may be due to their smooth formant trajectories. But one has to pay the price in terms of substantial disk usage, i.e., a seven fold increase in memory requirements for seven-frame distances compared to single frame distances, as we need to store fourteen frames for each units to compute spectral distance measures.

We have performed several ANOVA tests on single-frame spectral distances and data obtained from perceptual listening experiments. The observations from these tests are as follows:

- The four distance measures are significantly different from one another for a given spectral feature
- The effect of distance measure on the spectral distance is much stronger that that due to spectral feature
- The spectral distance relate to sentences or diphthongs, spectral features and distance measures
- The spectral distance do not relate to whether it is the first or second sentence in the given diphthong
- The listeners' scores differ significantly from one sentence or diphthong to another and also one listener to another but the listeners' F value is much lower than the spectral features F value and the distance measures F value
- The test stimuli are significantly different for a given diphthong

Chapter 4

Weighted Spectral Distances

4.1 Introduction

In the previous chapter we presented correlation coefficients computed between mean perceptual scores and various spectral distance measures based on single frame and multi-frame vectors of MFCCs, LSFs and MCA coefficients. We also presented the correlations for distances based on LPC spectra. It was observed that LSFs produce good correlations in some cases and MFCCs and MCA parameters in some others. Also, similar observations were seen across the various distance metrics. Hence, if we combine these speech parameterisations by some optimal weighting, they may produce more 1% significant correlations.

Investigation of the distance measures in the previous chapter indicated that we can weight them in the following types:

1. A weighted sum of different speech parameterisations for a single distance (absolute/Euclidean/Mahalanobis/K-L)
2. Weighting individual coefficients of a single speech parameterisation (MFCC/LSF/MCA)[1]
3. Weighting each frame distance in multi-frame distances

In section 4.2, we present the procedure for computing weights by using least squares linear regression on the training data in closed set and open set (K-fold cross valida-

[1] LPC spectra have not been included due to their poor performance in single frame distances

tion) experiments. Then we present correlations of mean perceptual listener scores with weighted distances obtained from the three different types (listed before) in section 4.3. Concluding remarks on these various methods are presented in section 4.4.

4.2 Weight computation

Consider the weighted distance, D_w, as shown in the equation below:

$$D_w = \sum_{i=1}^{N} w_i \, D_i(e_j, s_j) \tag{4.1}$$

where, w_i is weight on distance D_i between two feature vectors (e_j and s_j). To compute the weights, we solve the following equation:

$$\sum_{i=1}^{N} w_i \, D_i(e_j, s_j) = L_j \qquad \forall j \tag{4.2}$$

where, L_j is the mean listener rating, N is number of feature-types (MFCC, LSF, MCA) or the number of components in a feature vector or the number of frames in the case of multi-frame distances.

This set of equations is over-determined (more equations than unknowns) as $j > i$, where j is the index ($1 \leq j \leq M$) of a join in the perceptual experiment results and M is total number of stimuli in the perceptual experiment. Weights can be computed by solving the above equation as a linear regression problem, i.e., using the least squares method to fit a linear model of the form:

$$D\,\mathbf{w} = \mathbf{L} \tag{4.3}$$

where $\mathbf{w}$ is a weights vector $\{w_1, \cdots, w_N\}$, D is an M-by-N distance matrix and $\mathbf{L}$ is the column vector $\{L_1, \cdots, L_M\}$. We also tried a linear regression with intercept as shown below:

$$D\,\mathbf{w} + \mathbf{b} = \mathbf{L} \tag{4.4}$$

Solving the above equations is trivial; we used the MATLAB backslash (\) operator, which uses QR decomposition with pivoting.

We started with closed set experiments, where we obtained weights on distances using all of our perceptual data. Then we performed open set experiments by splitting data

into a *training set* and *test set* using a standard cross-validation procedure. We used the training set to compute weights by using a linear model of forms with (equation 4.4) and without intercept (equation 4.3). Then used these weights on test set and obtained weighted distances. Finally, computed the correlation between these weighted distances and the mean listener scores.

4.2.1 Cross-validation

To employ the cross-validation procedure, we treated our perceptual data corresponding to the 10 sentences in table 3.2 as 10 different sets. First, we used one data set as the *test set* and remaining 9 data sets as the *training set*. Then, we repeated this process for each of the possible choices (10 in this case), i.e. until each data set has been used as a test set. This method is known as the *leave-one-out* method (Bishop 1995). Figure 4.1 describes partitioning of training and test data sets, where test sets in each run are shown shaded.

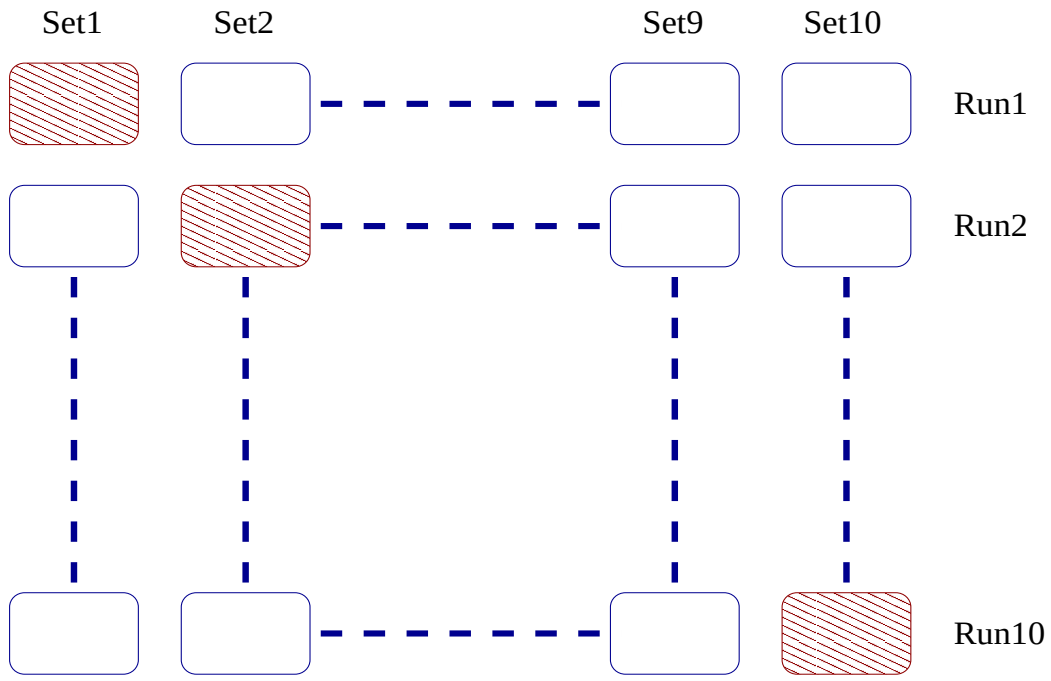

Figure 4.1: Schematic diagram of cross-validation procedure (adapted from Bishop (1995))

We also used another two ways of partitioning our full data. In one case, we used 2 data sets as test set and the remaining 8 data sets as training set. In the other case, 5 data sets were used as training set and the remaining 5 data sets as test set. Hence, we have three cases of open set experiments. We use the following terminology to represent these three cases in all our results throughout this thesis;

- ***open set(9+1)***: 9 data sets as training set and 1 data set as test set
- ***open set(8+2)***: 8 data sets as training set and 2 data sets as test set
- ***open set(5+5)***: 5 data sets as training set and 5 data sets as test set

We observed from initial correlation results that these are not good in some cases of *open set(8+2)* and *open set(5+5)*. The reason could be there is no data set from the same class[2] in training set corresponding to the same class of test set. For example, consider A, B, C, D and E are five different classes corresponding five diphthongs and in each class subscripts 1,2 represent the two data sets, i.e., A_1, A_2. In the case of *open set(8+2)*, the first run is:
$Training\ set : \{B_1, B_2, C_1, C_2, D_1, D_2, E_1, E_2\}$
$Test\ set : \{A_1, A_2\}$
Here, there is no training data corresponding to class A. Hence, we included one data set of class A in test set and one in training set. This produced a slight improvement in our results. In the case of *open set(5+5)* we achieved more improvement using one data set of each class in test set and in training set, i.e. using the following sequence in the first run:
$Training\ set : \{A_2, B_2, C_2, D_2, E_2\}$
$Test\ set : \{A_1, B_1, C_1, D_1, E_1\}$
And in the second run, we swapped the above two sets.

In the following sections we present the results for only the above discussed partitioning method for: *open set(8+2)* and *open set(5+5)*. But for *open set(9+1)* case, we use the leave-one-out method (Bishop 1995) for partitioning the data. In each case, we computed weights using the training set by the least squares linear regression method

[2] In our case each pair of sentences of one diphthong can be considered as *a class*

described in the previous section. Later, we used these weights to compute distances for the test set and thus to obtain correlations with mean listening scores.

4.3 Correlations with weighted distances

We carried out the three different types of weighting distance measures obtained from MFCC, LSF and MCA parameters, which are listed on page 79. In the next few sections, we present tables of correlation coefficients computed between mean listener scores and various weighted distances. We also present the global correlation coefficient which is the correlation between mean listener scores of all our ten sentences and corresponding weighted distances. As mentioned in section 3.3, we use the number of 1% significant correlations and global correlation coefficient to compare various weighted distances.

4.3.1 Weighted combinations of three speech parameterisations

Table 4.1 summarises correlations of absolute distances based on MFCCs, LSFs and MCA coefficients and their weighted sum with mean listener ratings. The correlation coefficients above the 1% significant level are highlighted. We present the results for closed set experiments, where weight training is performed on the full data, and various open set experiments (*open set(9+1)*, *open set(8+2)* and *open set(5+5)*). For each of these four cases we obtained weights using the least squares method by fitting the linear model without intercept (i.e., using equation 4.3) and with intercept (using equation 4.4). The correlation values listed in first column for each case were obtained using weights from linear model and those are in second column were computed using weights from linear model with intercept.

The three weights (normalised) obtained from linear model for absolute distances using MFCCs, LSFs and MCAs in closed set experiments for the second sentence of *ey* (i.e. second element in the first column of the table 4.1) are 0.15, 0.44, 0.41 respectively. From the correlation results, we can observe that there is no improvement in the absolute correlation values and in the number of 1% significant correlations compared to those obtained from individual absolute distances of MFCCs, LSFs and MCA coefficients, which are reported in the previous chapter (tables 3.4, 3.5 and 3.6).

Diphthong	*Closed set*		*Open set(9+1)*		*Open set(8+2)*		*Open set(5+5)*	
ey	0.28	0.16	0.27	0.13	0.28	0.20	0.20	0.12
	0.72	**0.75**	**0.72**	**0.74**	**0.72**	**0.73**	**0.69**	**0.67**
ow	0.29	0.37	0.29	0.22	0.30	0.24	0.35	0.35
	0.53	**0.55**	**0.53**	**0.54**	**0.53**	**0.51**	**0.54**	**0.57**
ay	0.27	0.27	0.26	0.23	0.25	0.20	0.29	0.10
	0.61	**0.59**	**0.60**	**0.58**	**0.60**	**0.51**	**0.61**	**0.62**
aw	0.36	0.20	0.34	0.06	0.33	0.00	0.26	0.07
	0.76	**0.77**	**0.76**	**0.73**	**0.76**	**0.71**	**0.74**	**0.71**
oy	0.11	0.07	0.11	0.06	0.11	0.01	0.08	0.11
	0.13	0.14	0.13	0.13	0.13	0.13	0.13	0.12

Table 4.1: Correlation between perceptual scores and weighted **absolute** distances of MFCCs, LSFs and MCAs obtained using weights from closed and open set experiments

The correlation results for absolute distances based on MFCCs, LSFs and MCAs with their respective delta coefficients showed in table 4.2. In this case the number of 1% significant correlations are more (**five**) than their individual counterparts (tables 3.4, 3.5 and 3.6) in all our open set and closed set experiments using linear regression with intercept. The same observation is made in the *closed set* experiment using linear regression without intercept.

It is clear from these results that we can improve correlations by setting weights on individual distances, e.g. MCAs produce good correlations for the first sentence in the *oy* diphthong pair (table 3.6), MFCCs and LSFs have good correlations for the second sentence in *ay* diphthong pair (tables 3.4 and 3.5). However, a weighted measure achieves good correlations in both cases (table 4.2). Also, it narrowly missed our 1% significant level (which is 0.446) for the first sentence in *aw* diphthong pair. If we consider this also, then this weighted distance resulted in **six** significant correlations.

The correlation coefficients of Euclidean distances of all the three spectral features and their weighted sum with mean listener ratings are reported in table 4.3. The weighted measure yielded slightly improved correlation results, **five** significant corre-

Diphthong	*Closed set*		*Open set(9+1)*		*Open set(8+2)*		*Open set(5+5)*	
ey	0.36	0.35	0.35	0.34	0.31	0.31	0.36	0.33
	0.66	**0.63**	**0.67**	**0.66**	**0.65**	**0.61**	**0.69**	**0.70**
ow	0.18	0.14	0.17	0.12	0.17	0.12	0.15	0.11
	0.56	**0.54**	**0.57**	**0.53**	**0.57**	**0.51**	**0.55**	**0.55**
ay	0.11	0.01	0.10	-0.03	0.11	-0.05	0.04	-0.08
	0.61	**0.52**	**0.58**	**0.47**	**0.57**	**0.45**	**0.69**	**0.70**
aw	0.42	0.44	0.42	0.44	0.43	0.43	0.44	0.41
	0.75	**0.72**	**0.75**	**0.69**	**0.75**	**0.71**	**0.77**	**0.78**
oy	0.45	**0.50**	0.45	**0.49**	0.46	**0.51**	**0.51**	**0.54**
	0.29	0.35	0.23	0.26	0.23	0.24	0.22	0.23

Table 4.2: Correlation between perceptual scores and weighted **absolute (with deltas)** distances MFCCs, LSFs and MCAs obtained using weights from closed and open set experiments

lations obtained for both *closed set* and *open set(5+5)* using linear regression with intercept. The results obtained from weighting the Euclidean distances computed from delta coefficients are showed in table 4.4. Here, also we achieved **five** significant correlations compared to four obtained using individual distances (tables 3.4, 3.5 and 3.6).

In the tables 4.5 and 4.6, we present correlations between perceptual scores and Mahalanobis distances of MFCCs, LSFs and MCA coefficients along with their deltas. Unfortunately, these results are worse than individual distances, where we observed **six** 1% significant correlations in the case of Mahalanobis distances computed on LSFs and their delta coefficients (shown in table 3.5).

Diphthong	*Closed set*		*Open set(9+1)*		*Open set(8+2)*		*Open set(5+5)*	
ey	0.26	0.05	0.26	0.00	0.26	0.11	0.14	-0.04
	0.70	**0.73**	**0.68**	**0.72**	**0.68**	**0.72**	**0.67**	**0.67**
ow	0.29	**0.47**	0.28	0.18	0.30	0.20	0.42	**0.49**
	0.55	**0.59**	**0.55**	**0.58**	**0.54**	**0.56**	**0.55**	**0.60**
ay	0.27	0.28	0.26	0.23	0.25	0.21	0.34	0.20
	0.60	**0.55**	**0.60**	**0.53**	**0.60**	**0.45**	**0.60**	**0.60**
aw	0.38	0.14	0.36	0.03	0.36	-0.03	0.20	0.06
	0.78	**0.78**	**0.78**	**0.75**	**0.78**	**0.73**	**0.77**	**0.76**
oy	0.13	0.09	0.13	0.06	0.13	-0.00	0.06	0.12
	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14

Table 4.3: Correlation between perceptual scores and weighted **Euclidean** distances MFCCs, LSFs and MCAs obtained using weights from closed and open set experiments

Diphthong	*Closed set*		*Open set(9+1)*		*Open set(8+2)*		*Open set(5+5)*	
ey	0.28	0.26	0.27	0.23	0.27	0.25	0.27	0.22
	0.65	**0.64**	**0.66**	**0.67**	**0.65**	**0.62**	**0.68**	**0.66**
ow	0.22	0.19	0.20	0.15	0.21	0.17	0.18	0.14
	0.61	**0.60**	**0.61**	**0.60**	**0.61**	**0.58**	**0.60**	**0.59**
ay	0.18	0.12	0.17	0.08	0.17	0.07	0.12	0.00
	0.60	**0.56**	**0.58**	**0.53**	**0.57**	**0.51**	**0.67**	**0.65**
aw	0.41	0.42	0.41	0.42	0.41	0.41	0.41	0.37
	0.78	**0.76**	**0.78**	**0.74**	**0.77**	**0.76**	**0.79**	**0.78**
oy	0.46	**0.50**	0.46	**0.48**	0.47	**0.48**	**0.51**	**0.54**
	0.25	0.28	0.22	0.23	0.22	0.22	0.21	0.18

Table 4.4: Correlation between perceptual scores and weighted **Euclidean (with deltas)** distances MFCCs, LSFs and MCAs obtained using weights from closed and open set experiments

Diphthong	*Closed set*		*Open set(9+1)*		*Open set(8+2)*		*Open set(5+5)*	
ey	0.26	0.22	0.25	0.20	0.23	0.18	0.26	0.24
	0.71	**0.73**	**0.71**	**0.72**	**0.68**	**0.63**	**0.70**	**0.75**
ow	0.30	0.34	0.28	0.27	0.28	0.25	0.30	0.30
	0.57	**0.53**	**0.58**	**0.51**	**0.57**	**0.48**	**0.59**	**0.55**
ay	0.28	0.21	0.26	0.17	0.26	0.15	0.16	0.08
	0.56	**0.54**	**0.55**	**0.54**	**0.55**	**0.51**	**0.56**	**0.55**
aw	0.44	0.37	0.43	0.30	0.42	0.28	0.42	0.36
	0.80	**0.81**	**0.80**	**0.81**	**0.80**	**0.81**	**0.79**	**0.81**
oy	0.30	0.31	0.30	0.31	0.30	0.30	0.32	0.30
	0.21	0.23	0.20	0.20	0.19	0.19	0.19	0.20

Table 4.5: Correlation between perceptual scores and weighted **Mahalanobis** distances MFCCs, LSFs and MCAs obtained using weights from closed and open set experiments

Diphthong	*Closed set*		*Open set(9+1)*		*Open set(8+2)*		*Open set(5+5)*	
ey	0.36	0.36	0.36	0.36	0.33	0.33	0.37	0.34
	0.63	**0.61**	**0.63**	**0.60**	**0.60**	**0.51**	**0.64**	**0.62**
ow	0.12	0.08	0.07	0.01	0.07	0.01	0.05	0.00
	0.59	**0.58**	**0.59**	**0.58**	**0.59**	**0.57**	**0.59**	**0.54**
ay	0.17	0.09	0.16	0.07	0.17	0.05	0.09	-0.00
	0.66	**0.65**	**0.66**	**0.62**	**0.65**	**0.62**	**0.62**	**0.52**
aw	**0.48**	**0.58**	**0.46**	**0.52**	**0.46**	**0.49**	**0.57**	**0.56**
	0.80	**0.80**	**0.80**	**0.79**	**0.80**	**0.80**	**0.79**	**0.76**
oy	0.30	0.35	0.30	0.34	0.31	0.34	0.38	0.40
	0.33	0.41	0.30	0.33	0.29	0.32	0.31	0.40

Table 4.6: Correlation between perceptual scores and weighted **Mahalanobis (with deltas)** distances MFCCs, LSFs and MCAs obtained using weights from closed and open set experiments

We also computed correlation coefficients in the case of weighted K-L distance, where we weighted LSF and MCA K-L distances. These are presented in table 4.7. But we did not observe any improvement over individual LSF or MCA K-L distances.

Diphthong	*Closed set*		*Open set(9+1)*		*Open set(8+2)*		*Open set(5+5)*	
ey	0.32	0.26	0.32	0.24	0.33	0.27	0.31	0.31
	0.73	**0.72**	**0.73**	**0.72**	**0.73**	**0.73**	**0.73**	**0.70**
ow	0.31	0.36	0.29	0.28	0.31	0.30	0.32	0.32
	0.44	**0.45**	0.43	**0.45**	0.43	**0.45**	0.44	0.44
ay	0.36	0.40	0.36	0.39	0.35	0.39	0.37	0.37
	0.63	**0.63**	**0.63**	**0.62**	**0.63**	**0.61**	**0.63**	**0.60**
aw	0.20	0.11	0.19	0.05	0.19	0.03	0.18	0.19
	0.80	**0.78**	**0.80**	**0.76**	**0.80**	**0.75**	**0.80**	**0.74**
oy	0.16	0.14	0.16	0.14	0.16	0.12	0.16	0.16
	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.31

Table 4.7: Correlation between perceptual scores and weighted **Kullback-Leibler (K-L)** distances LSFs and MCAs obtained using weights from closed and open set experiments

Finally, in figure 4.2 we compare all the cases discussed in this section, using global correlation coefficients and number of 1% significant correlations. The bars in each pair represent the results using weights obtained from the least squares linear regression without intercept and with intercept respectively. The numbers on top of each bar are the number of 1% significant correlations. The global correlation coefficient is calculated using perceptual scores of all our stimuli and corresponding distance measures. The closed set experiments resulted in high global correlation coefficients. The weights computed using linear regression produced higher global correlations compared to those obtained from linear regression with intercept.

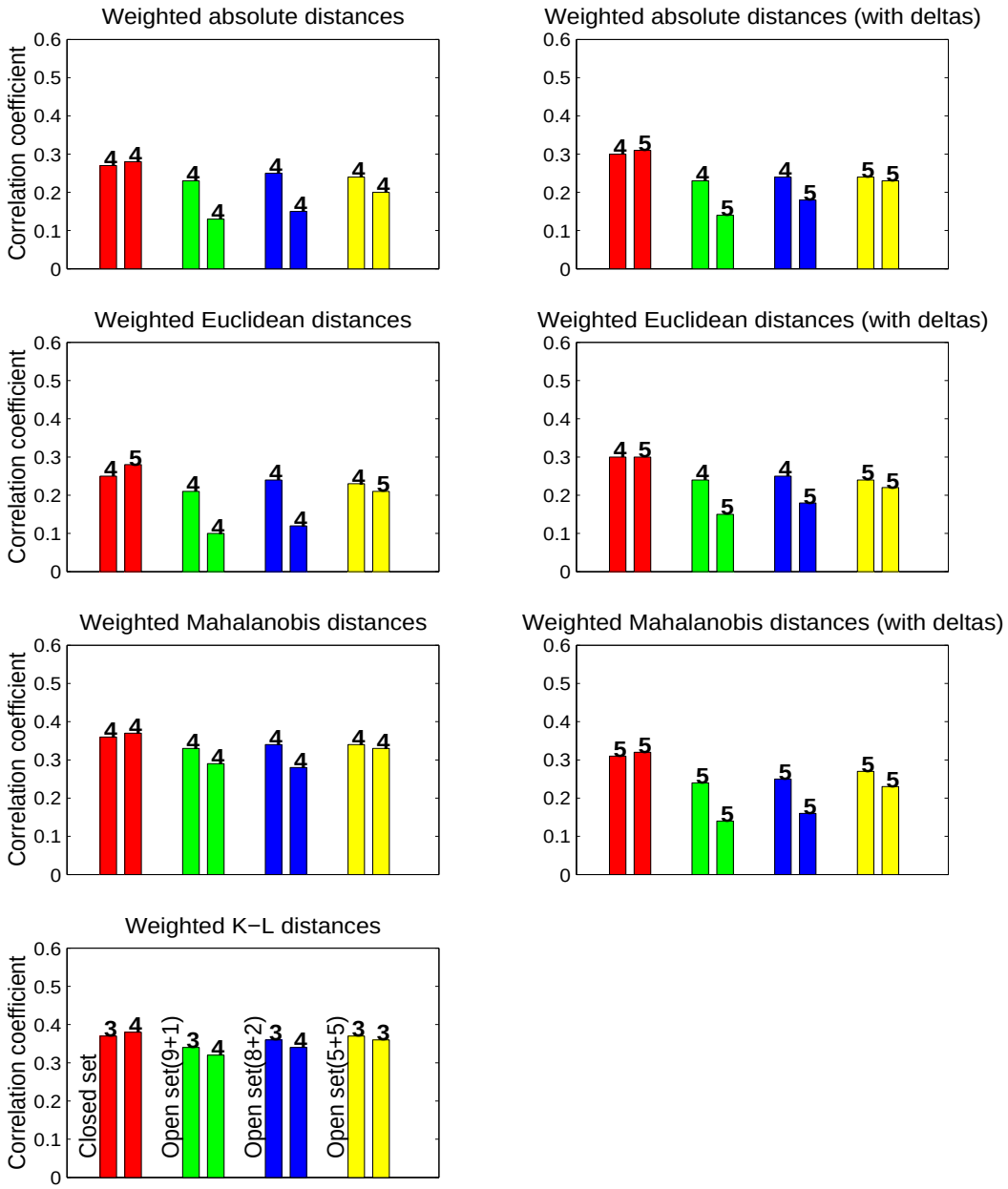

Figure 4.2: Comparison of global correlation coefficients among different weighted **distance metrics** for *closed set* and three open set experiments (*open set(9+1)*, *open set(8+2)* and *open set(5+5)*). The numbers on top of each bar are number of 1% significant correlations.

4.3.2 Weights on individual features of speech parameterisations

In order to improve correlations further, we constructed distance measures by applying weights on individual coefficients of three parameterisations used, MFCCs, LSFs and MCAs. The following tables present the correlation coefficients obtained between mean listener scores and the absolute distances computed on weighted individual parameters of MFCCs, LSFs and MCAs with and without delta coefficients.

Diphthong	*Closed set*		*Open set(9+1)*		*Open set(8+2)*		*Open set(5+5)*	
ey	0.22	0.30	0.21	0.26	0.23	0.32	0.24	0.37
	0.72	**0.63**	**0.71**	**0.56**	**0.72**	**0.59**	**0.75**	**0.63**
ow	0.21	0.31	0.18	0.18	0.18	0.17	0.18	0.18
	0.47	0.38	**0.46**	0.15	0.44	0.05	0.41	0.17
ay	**0.52**	**0.58**	**0.57**	**0.54**	**0.53**	**0.43**	**0.56**	**0.59**
	0.59	**0.60**	**0.55**	**0.46**	**0.51**	0.32	**0.57**	0.38
aw	0.36	0.36	0.30	0.00	0.29	-0.14	0.40	0.23
	0.77	**0.77**	**0.78**	**0.75**	**0.79**	**0.74**	**0.73**	**0.63**
oy	0.30	0.45	0.27	0.24	0.02	-0.12	0.11	-0.20
	0.18	0.30	0.12	0.18	0.11	0.16	0.13	0.21

Table 4.8: Correlation between perceptual scores and weighted individual **MFCCs** absolute distances obtained using weights from closed and open set experiments

Overall we observed more significant correlations with delta coefficients. We achieved our highest (**eight**) 1% significant correlations in the closed set experiment of weighted MFCCs and LSFs as reported in tables 4.9 and 4.11. However, in open set experiments we achieved fewer significant correlations, at the most **five** for the LSFs with deltas in *open set(5+5)* (table 4.11). Hence, we can not generalise these results.

Diphthong	*Closed set*		*Open set(9+1)*		*Open set(8+2)*		*Open set(5+5)*	
ey	0.30	0.32	0.27	0.24	0.29	0.29	0.36	0.39
	0.70	**0.64**	**0.68**	**0.53**	**0.70**	**0.60**	**0.71**	**0.66**
ow	0.35	**0.53**	0.30	0.40	0.32	0.39	0.28	0.33
	0.44	0.43	0.41	0.26	0.35	0.19	0.27	0.21
ay	0.39	**0.45**	0.39	0.36	0.33	0.27	0.25	0.23
	0.62	**0.68**	**0.54**	**0.52**	**0.50**	**0.46**	**0.57**	0.33
aw	0.31	**0.48**	0.16	0.12	0.16	0.11	0.32	0.35
	0.76	**0.73**	**0.74**	**0.67**	**0.77**	**0.68**	**0.75**	**0.60**
oy	**0.48**	**0.52**	0.34	0.27	0.17	0.13	0.04	-0.11
	0.50	**0.52**	0.27	0.22	0.30	0.24	0.33	0.11

Table 4.9: Correlation between perceptual scores and weighted individual **MFCCs (with detlas)** absolute distances obtained using weights from closed and open set experiments

Diphthong	*Closed set*		*Open set(9+1)*		*Open set(8+2)*		*Open set(5+5)*	
ey	0.16	0.16	0.09	0.00	0.07	-0.02	0.15	-0.06
	0.74	**0.81**	**0.72**	**0.77**	**0.67**	**0.70**	**0.79**	**0.70**
ow	0.25	0.06	0.18	-0.06	0.15	-0.08	0.17	0.11
	0.38	0.27	0.30	-0.17	0.22	-0.21	0.25	0.02
ay	0.32	**0.45**	0.28	0.35	0.28	0.38	0.30	0.28
	0.54	0.37	**0.53**	0.24	**0.53**	0.29	**0.52**	0.24
aw	0.25	0.35	0.22	0.19	0.23	0.19	0.26	0.15
	0.77	**0.72**	**0.76**	**0.68**	**0.76**	**0.72**	**0.75**	**0.65**
oy	0.13	0.35	0.08	0.20	0.04	0.17	0.01	0.06
	0.53	**0.58**	0.39	0.41	0.36	0.38	**0.45**	0.42

Table 4.10: Correlation between perceptual scores and weighted individual **LSFs** absolute distances obtained using weights from closed and open set experiments

Diphthong	*Closed set*		*Open set(9+1)*		*Open set(8+2)*		*Open set(5+5)*	
ey	0.43	**0.53**	0.28	0.20	0.22	0.14	0.24	0.17
	0.73	**0.79**	**0.70**	**0.71**	**0.64**	**0.66**	**0.74**	**0.69**
ow	0.01	0.03	-0.12	-0.22	-0.12	-0.18	-0.15	-0.13
	0.16	0.12	-0.14	-0.39	-0.10	-0.36	-0.08	-0.19
ay	0.36	**0.46**	0.30	0.28	0.23	0.21	**0.49**	**0.43**
	0.69	**0.53**	**0.61**	0.36	**0.62**	0.36	**0.67**	0.16
aw	**0.46**	**0.54**	0.37	0.15	0.38	0.13	**0.52**	-0.01
	0.77	**0.72**	**0.75**	**0.67**	**0.76**	**0.72**	**0.75**	0.42
oy	0.39	**0.53**	0.26	0.26	0.22	0.19	0.21	0.08
	0.62	**0.48**	0.24	-0.03	0.21	-0.05	0.20	-0.10

Table 4.11: Correlation between perceptual scores and weighted individual **LSFs (with detlas)** absolute distances obtained using weights from closed and open set experiments

Diphthong	*Closed set*		*Open set(9+1)*		*Open set(8+2)*		*Open set(5+5)*	
ey	0.34	0.24	0.34	0.23	0.34	0.25	0.35	0.33
	0.57	0.42	**0.56**	0.38	**0.56**	0.40	**0.56**	**0.47**
ow	0.12	0.38	-0.00	0.10	-0.02	0.08	-0.01	0.03
	0.52	**0.66**	**0.52**	**0.65**	0.43	**0.59**	0.43	**0.55**
ay	0.08	0.15	0.09	0.14	0.10	0.16	0.05	0.07
	0.27	0.16	0.26	0.11	0.23	0.12	0.29	0.09
aw	0.44	**0.57**	**0.45**	**0.50**	0.44	**0.53**	**0.46**	**0.59**
	0.74	0.37	**0.73**	0.07	**0.72**	0.06	**0.71**	0.38
oy	0.10	0.10	0.08	-0.01	-0.01	-0.10	0.09	-0.13
	0.24	0.24	0.15	0.17	0.15	0.15	0.17	0.15

Table 4.12: Correlation between perceptual scores and weighted individual **MCAs** absolute distances obtained using weights from closed and open set experiments

Diphthong	*Closed set*		*Open set(9+1)*		*Open set(8+2)*		*Open set(5+5)*	
ey	0.40	0.39	0.39	0.34	0.38	0.30	0.39	0.34
	0.62	**0.65**	**0.59**	**0.55**	**0.60**	**0.56**	**0.62**	**0.56**
ow	0.16	0.43	-0.04	0.01	-0.04	0.01	-0.06	-0.09
	0.63	**0.68**	**0.64**	**0.67**	**0.59**	**0.67**	**0.56**	**0.64**
ay	0.22	0.22	0.21	0.17	0.21	0.17	0.17	0.18
	0.24	0.14	0.22	0.10	0.21	0.11	0.24	0.03
aw	**0.64**	**0.68**	**0.64**	**0.59**	**0.64**	**0.60**	**0.66**	**0.65**
	0.63	0.41	**0.61**	0.24	**0.62**	0.19	**0.64**	0.33
oy	**0.51**	0.39	0.47	0.20	0.44	0.13	0.39	0.03
	0.39	0.40	0.25	0.29	0.25	0.27	0.27	0.27

Table 4.13: Correlation between perceptual scores and weighted individual **MCAs (with detlas)** absolute distances obtained using weights from closed and open set experiments

In figure 4.3 we compare the various cases discussed in this section, using global correlation coefficients and number of 1% significant correlations. Here also closed set experiments resulted in high global correlation coefficients. In the case of MFCC and LSF, we achieved **eight** 1% significant correlations. Though the weights computed using linear regression produced higher global correlations in more number of cases than those obtained from linear regression with intercept, the latter has a greater number (eight) of 1% significant correlations.

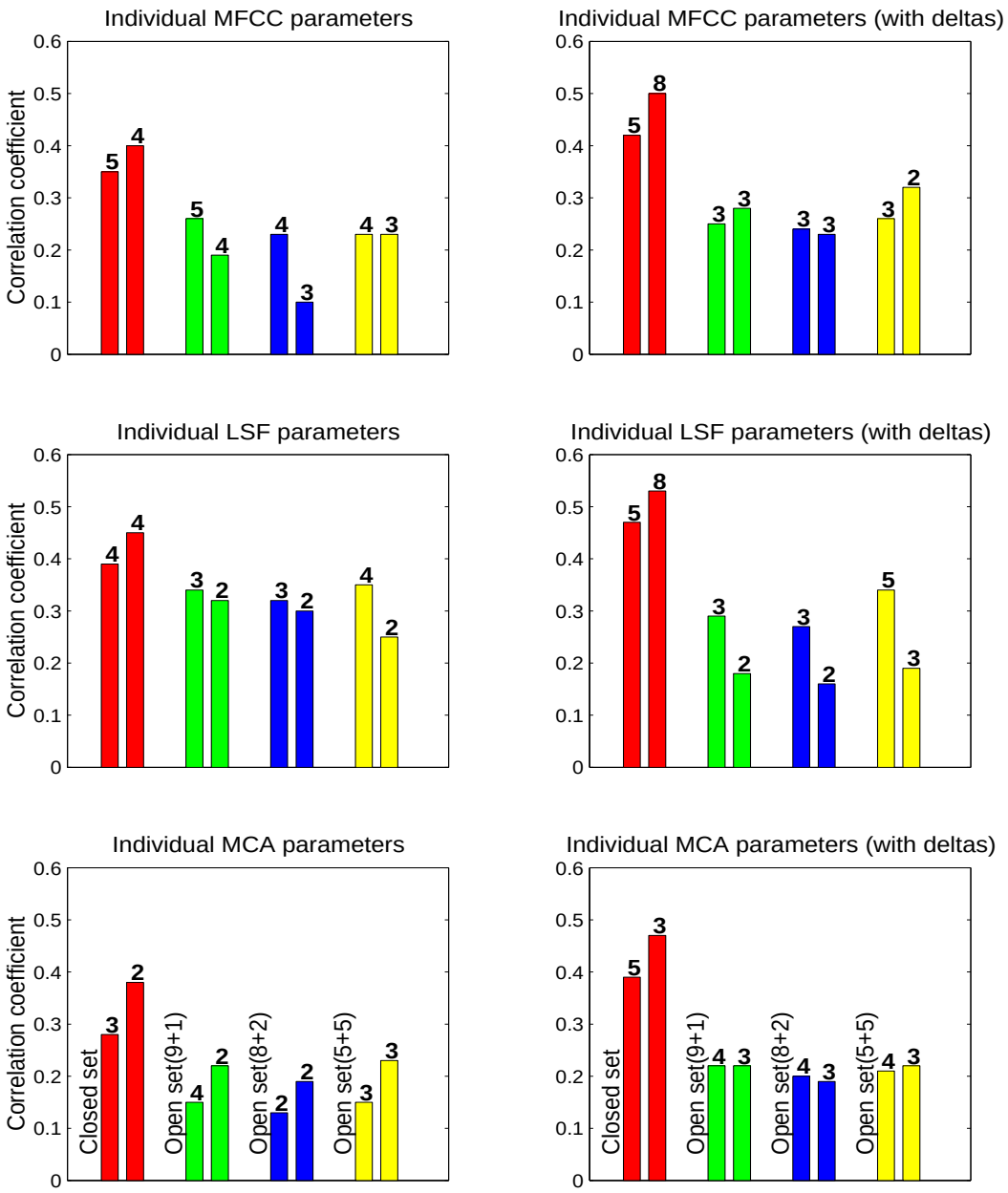

Figure 4.3: Comparison of global correlation coefficients among different weighted **individual spectral features** for distance computation for *closed set* and three open set experiments (*open set(9+1)*, *open set(8+2)* and *open set(5+5)*). The numbers on top of the bar are number of 1% significant correlations.

4.3.3 Weighted multi-frame distances

4.3.3.1 Using triangular weights

We carried out weighting on each frame distance in our multi-frame distances, presented in 3.3.2. First we started with triangular weighting, i.e. giving highest weight to the middle frame and the lower weights to the frames away from the join, as shown in the figure 4.4. The triangular weighting can be described using the equation below:

$$w_i = \begin{cases} \frac{2i}{N+1} & 1 \leq i \leq \frac{N+1}{2} \\ \frac{2(N-i+1)}{N+1} & \frac{N+1}{2} \leq i \leq N \end{cases} \tag{4.5}$$

where N is number of frames, in our case it is odd (3 or 5 or 7 or 9).

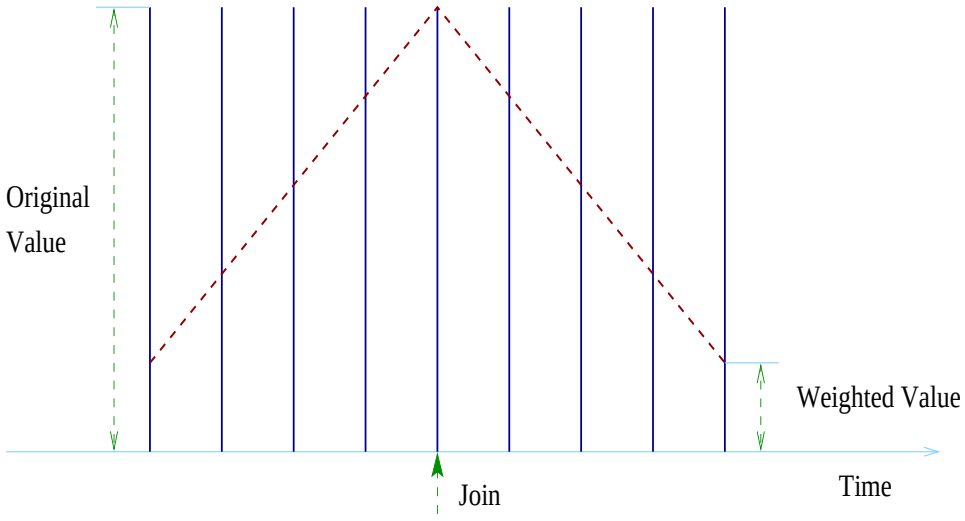

Figure 4.4: Triangular weighting

The correlation results for various frame distances using MFCC, LSF and MCA parameters are presented in the following tables. In the case of MFCCs, there is some improvement in the number of significant correlations compared to average multi-frame distances presented in section 3.3.2. Mainly, the correlation values at the boundaries of significance level were boosted up, for example in table 4.14, the fifth column of the first row of *ay*, the correlation value improved and crossed the significance level. In the case of LSFs and MCAs there are also a few more 1% significant correlations compared to average multi-frame distances presented in the previous chapter (section

3.3.2). However, we do not consider it as major improvement as it does not yield more than **seven** 1% significant correlations in any of these cases.

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	0.29	0.31	0.28	0.33	0.20	0.32
	0.71	**0.70**	**0.73**	**0.71**	**0.77**	**0.69**
ow	0.28	0.26	0.27	0.22	0.28	0.19
	0.57	**0.57**	**0.55**	**0.54**	**0.58**	**0.57**
ay	0.32	0.31	0.33	0.26	**0.42**	0.24
	0.67	**0.69**	**0.66**	**0.74**	**0.62**	**0.71**
aw	0.32	0.33	0.32	0.32	0.23	0.27
	0.78	**0.79**	**0.78**	**0.79**	**0.79**	**0.80**
oy	0.06	0.07	0.06	0.11	0.22	0.23
	0.16	0.19	0.15	0.20	0.19	0.28

Table 4.14: Correlation between perceptual scores and **triangular** weighted distance measures computed using **three** frames of **MFCC**s

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	0.36	0.38	0.35	0.39	0.27	0.37
	0.72	**0.71**	**0.73**	**0.71**	**0.77**	**0.69**
ow	0.27	0.26	0.26	0.23	0.27	0.20
	0.57	**0.57**	**0.54**	**0.55**	**0.58**	**0.58**
ay	0.33	0.31	0.33	0.27	**0.42**	0.27
	0.70	**0.73**	**0.69**	**0.76**	**0.64**	**0.73**
aw	0.29	0.31	0.29	0.33	0.20	0.31
	0.80	**0.80**	**0.79**	**0.80**	**0.80**	**0.79**
oy	0.03	0.06	0.05	0.11	0.18	0.24
	0.18	0.21	0.16	0.21	0.21	0.30

Table 4.15: Correlation between perceptual scores and **triangular** weighted distance measures computed using **five** frames of **MFCC**s

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	0.38	0.40	0.36	0.40	0.28	0.38
	0.73	**0.72**	**0.73**	**0.71**	**0.77**	**0.69**
ow	0.27	0.25	0.25	0.22	0.27	0.20
	0.58	**0.59**	**0.56**	**0.57**	**0.59**	**0.59**
ay	0.33	0.32	0.33	0.27	**0.42**	0.27
	0.72	**0.75**	**0.71**	**0.77**	**0.65**	**0.73**
aw	0.27	0.29	0.27	0.31	0.18	0.30
	0.81	**0.81**	**0.81**	**0.80**	**0.81**	**0.79**
oy	0.04	0.09	0.05	0.13	0.17	0.29
	0.20	0.22	0.17	0.21	0.22	0.29

Table 4.16: Correlation between perceptual scores and **triangular** weighted distance measures computed using **seven** frames of **MFCC**s

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	0.39	0.41	0.37	0.41	0.30	0.39
	0.74	**0.72**	**0.74**	**0.71**	**0.77**	**0.69**
ow	0.27	0.25	0.25	0.23	0.27	0.21
	0.58	**0.58**	**0.57**	**0.57**	**0.59**	**0.59**
ay	0.34	0.32	0.33	0.28	**0.42**	0.27
	0.73	**0.76**	**0.72**	**0.77**	**0.66**	**0.74**
aw	0.25	0.26	0.25	0.28	0.17	0.26
	0.82	**0.82**	**0.82**	**0.81**	**0.82**	**0.79**
oy	0.09	0.13	0.09	0.17	0.19	0.32
	0.20	0.22	0.19	0.22	0.23	0.29

Table 4.17: Correlation between perceptual scores and **triangular** weighted distance measures computed using **nine** frames of **MFCC**s

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*		*K-L*
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.04	0.14	0.12	0.26	0.13	0.35	0.29
	0.75	**0.73**	**0.76**	**0.74**	**0.74**	**0.62**	**0.74**
ow	0.36	0.24	0.33	0.20	0.38	0.16	0.30
	0.53	**0.51**	**0.47**	**0.45**	0.43	0.35	**0.47**
ay	0.14	0.06	0.09	0.00	0.13	0.01	0.37
	0.51	**0.67**	**0.52**	**0.71**	**0.48**	**0.54**	**0.66**
aw	0.23	0.38	0.13	0.35	0.26	**0.59**	0.17
	0.80	**0.79**	**0.79**	**0.77**	**0.79**	**0.67**	**0.80**
oy	0.21	0.24	0.16	0.20	0.33	0.39	0.17
	0.14	0.28	0.17	0.29	0.22	0.44	0.32

Table 4.18: Correlation between perceptual scores and **triangular** weighted distance measures computed using **three** frames of **LSF**s

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*		*K-L*
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.04	0.16	0.13	0.27	0.14	0.36	0.30
	0.75	**0.72**	**0.77**	**0.73**	**0.75**	**0.62**	**0.74**
ow	0.34	0.23	0.31	0.20	0.36	0.17	0.28
	0.53	**0.50**	**0.46**	**0.46**	**0.45**	0.38	**0.47**
ay	0.12	0.06	0.08	0.01	0.11	0.02	0.37
	0.55	**0.71**	**0.55**	**0.74**	**0.52**	**0.61**	**0.69**
aw	0.24	0.40	0.15	0.38	0.29	**0.60**	0.16
	0.80	**0.79**	**0.80**	**0.78**	**0.79**	**0.70**	**0.80**
oy	0.23	0.29	0.19	0.29	0.34	**0.48**	0.19
	0.15	0.26	0.18	0.26	0.22	0.38	0.33

Table 4.19: Correlation between perceptual scores and **triangular** weighted distance measures computed using **five** frames of **LSF**s

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*		*K-L*
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.05	0.18	0.13	0.28	0.15	0.38	0.32
	0.75	**0.72**	**0.77**	**0.73**	**0.75**	**0.63**	**0.73**
ow	0.32	0.22	0.29	0.20	0.35	0.17	0.27
	0.52	**0.49**	**0.46**	**0.46**	**0.47**	0.39	**0.46**
ay	0.11	0.05	0.07	0.01	0.09	0.01	0.37
	0.58	**0.73**	**0.57**	**0.76**	**0.54**	**0.65**	**0.70**
aw	0.25	0.42	0.16	0.39	0.30	**0.61**	0.16
	0.81	**0.80**	**0.81**	**0.80**	**0.80**	**0.72**	**0.80**
oy	0.24	0.31	0.22	0.34	0.36	**0.49**	0.21
	0.16	0.25	0.18	0.25	0.21	0.35	0.34

Table 4.20: Correlation between perceptual scores and **triangular** weighted distance measures computed using **seven** frames of **LSF**s

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*		*K-L*
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.06	0.20	0.14	0.29	0.17	0.39	0.33
	0.74	**0.71**	**0.77**	**0.72**	**0.75**	**0.61**	**0.73**
ow	0.33	0.23	0.30	0.22	0.35	0.19	0.27
	0.51	**0.48**	0.45	**0.45**	**0.47**	0.38	**0.45**
ay	0.10	0.05	0.07	0.01	0.07	0.00	0.37
	0.59	**0.74**	**0.59**	**0.76**	**0.56**	**0.67**	**0.71**
aw	0.26	0.40	0.17	0.37	0.31	**0.57**	0.16
	0.82	**0.81**	**0.82**	**0.81**	**0.81**	**0.73**	**0.81**
oy	0.26	0.29	0.23	0.33	0.38	0.41	0.22
	0.15	0.24	0.17	0.24	0.19	0.34	0.33

Table 4.21: Correlation between perceptual scores and **triangular** weighted distance measures computed using **nine** frames of **LSF**s

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*		*K-L*
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.40	0.25	0.38	0.31	0.34	0.35	**0.48**
	0.65	**0.48**	**0.66**	**0.50**	0.44	**0.48**	**0.65**
ow	0.25	0.20	0.28	0.25	0.28	0.22	0.22
	0.46	**0.48**	**0.45**	**0.53**	**0.55**	**0.57**	0.39
ay	-0.03	0.07	-0.04	0.02	0.03	0.05	0.12
	0.47	0.22	0.43	0.25	**0.45**	**0.54**	**0.52**
aw	**0.47**	0.24	0.38	0.32	**0.56**	**0.61**	0.39
	0.76	**0.52**	**0.78**	**0.53**	**0.74**	**0.74**	**0.80**
oy	0.36	**0.50**	0.33	**0.55**	0.36	0.36	0.23
	0.15	0.32	0.18	**0.47**	0.32	**0.74**	0.31

Table 4.22: Correlation between perceptual scores and **triangular** weighted distance measures computed using **three** frames of **MCA**s

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*		*K-L*
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.39	0.24	0.39	0.31	0.31	0.34	**0.47**
	0.68	**0.49**	**0.71**	**0.51**	**0.49**	**0.50**	**0.66**
ow	0.30	0.20	0.33	0.26	0.31	0.27	0.25
	0.50	**0.48**	**0.47**	**0.53**	**0.56**	**0.59**	0.41
ay	-0.03	0.02	-0.05	-0.02	0.02	0.04	0.12
	0.51	0.23	**0.46**	0.25	**0.49**	**0.55**	**0.57**
aw	0.43	0.22	0.35	0.28	**0.54**	**0.60**	0.35
	0.76	**0.52**	**0.79**	**0.53**	**0.73**	**0.73**	**0.80**
oy	0.41	**0.51**	0.38	**0.55**	0.43	0.41	0.28
	0.17	0.38	0.19	**0.50**	0.34	**0.77**	0.33

Table 4.23: Correlation between perceptual scores and **triangular** weighted distance measures computed using **five** frames of **MCA**s

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*		*K-L*
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.38	0.23	0.39	0.31	0.28	0.35	**0.47**
	0.68	**0.49**	**0.72**	**0.52**	**0.52**	**0.52**	**0.66**
ow	0.33	0.21	0.36	0.26	0.32	0.28	0.28
	0.53	**0.53**	**0.49**	**0.57**	**0.58**	**0.59**	**0.45**
ay	-0.04	-0.00	-0.05	-0.04	0.01	0.03	0.12
	0.54	0.24	**0.50**	0.26	**0.51**	**0.57**	**0.60**
aw	0.41	0.22	0.34	0.27	**0.52**	**0.56**	0.32
	0.75	**0.52**	**0.77**	**0.54**	**0.72**	**0.71**	**0.79**
oy	**0.50**	**0.49**	0.47	**0.54**	**0.51**	0.34	0.33
	0.27	0.44	0.29	**0.55**	0.42	**0.77**	0.37

Table 4.24: Correlation between perceptual scores and **triangular** weighted distance measures computed using **seven** frames of **MCA**s

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*		*K-L*
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.32	0.21	0.34	0.28	0.24	0.32	0.44
	0.68	**0.50**	**0.72**	**0.53**	**0.54**	**0.52**	**0.65**
ow	0.37	0.24	0.40	0.29	0.35	0.32	0.30
	0.54	**0.52**	**0.50**	**0.55**	**0.57**	**0.53**	**0.46**
ay	-0.03	0.00	-0.05	-0.03	0.00	0.02	0.13
	0.57	0.24	**0.53**	0.27	**0.54**	**0.60**	**0.63**
aw	0.39	0.22	0.33	0.26	**0.52**	**0.51**	0.30
	0.75	**0.53**	**0.77**	**0.54**	**0.72**	**0.72**	**0.78**
oy	**0.49**	**0.48**	**0.47**	**0.54**	**0.54**	0.33	0.34
	0.36	**0.49**	0.38	**0.58**	**0.47**	**0.76**	0.41

Table 4.25: Correlation between perceptual scores and **triangular** weighted distance measures computed using **nine** frames of **MCA**s

4.3.3.2 Using weights obtained from least squares method

We also computed weights using our least square method, presented in section 4.2. We report the correlation results for various (three, five, seven and nine) frame distances in Appendix B. These results contain three spectral parameterisations (MFCC, LSF and MCA coefficients), distances (Euclidean, Absolute, Mahalanobis and Kullback-Leibler) and our four experiment cases (closed set,*open set(9+1)*, *open set(8+2)* and *open set(5+5)*). In each experiment, for computing weights we fit the linear models with and without intercept. Correlations of weighted distances and mean listener scores were obtained for all of the above cases. We have not observed any improvement in number of significant correlations compared to average multi-frame distances presented in the previous chapter (section 3.3.2).

4.4 Conclusions

In this chapter, we discussed several weighted distance measures derived from three speech parameterisations (MFCC,LSF and MCA), distances (absolute, Euclidean, Mahalanobis and K-L) and various multi-frame (three, five, seven and nine) distances. The weights were computed using the standard least squares method by fitting the linear model with and without intercept. We carried out both closed set and open set experiments for computing weights using the least squares method. In open set experiments, training and test sets were chosen using a cross-validation procedure. We have also used triangular weighting for multi-frame distances, i.e. providing high weights to frames at the join and low weights to frames away from the join.

All the results presented in this chapter and in Appendix B are estimates of correlations from a model rather than measured correlations from experiments. By weighting distances of MFCC, LSF and MCA coefficients, we have improved the number of 1% significant correlations in some cases, but did not observe any significant difference among the correlation values. Hence, this improvement is marginal and considering the increased storage required for three parameterisations of the speech signal, these results are considered no better than our single frame distance, discussed in section 3.3.1.

We have achieved **eight** (our highest number) 1% significant correlations for distances computed by weighting individual MFCCs and their deltas (table 4.9) and also in the case of LSFs plus delta coefficients. Unfortunately, these results can not be generalised as weights for the above cases were computed in a closed set experiment. In open set experiments and other cases we have not achieved any improvement compared to individual distances.

The triangular weighted multi-frame distances produced more 1% significant correlations compared to their least squares counterpart. The former has better correlations than the average multi-frame distances in some cases.

Chapter 5

Kalman Filter-based Join Cost

5.1 Introduction

In this chapter, we propose the use of a **learned underlying representation** of the speech signal to define a join cost in unit selection speech synthesis. An appropriate underlying representation might be an articulatory one, so that it could give a useful model of articulation and provide better join cost computations as well as smoothing between two speech segments. However, it is very difficult to obtain articulatory data and therefore, we try to learn an underlying representation from the speech signal alone. In this case the representation is hidden and its behaviour is smooth and continuous as will be described in future sections.

We have chosen **linear dynamical models** (sometimes known as **Kalman filters**) to learn an underlying representation from the data (observations), which in our case will be line spectral frequencies (LSF)[1] of the natural speech. The linear dynamical model (LDM) is a probabilistic, continuous, state-space model which can also operate as an optimal smoothing filter. The LDM uses an underlying, hidden state space in which it describes smooth, continuous motion. This state space can be seen as an analogy for underlying articulator movement. Once the models are trained, they can infer the most likely (predicted) observations given an actual, noisy observation sequence. Then, we

[1] We preferred LSFs due to their interpolation properties and other spectral characteristics, discussed in section 2.3.2.1

can easily measure the error between actual and model predicted observations and thus derive a meaningful and theoretically motivated join cost function.

In section 5.2, we provide an introduction to linear dynamic models (LDM) and their properties. Then we discuss how to derive a join cost function using these models in section 5.3. In section 5.4, we report correlations between this new measure and mean listener scores obtained from our perceptual listening experiments, discussed in section 3.2. We present these results for various state dimensions, three different initialisations for the models and several analytical measures for join costs.

5.2 Linear dynamical models

Our presentation of linear dynamic models closely follows Frankel (2003). Linear dynamic models belong to class of state-space models, which have found many applications in various domains such as control systems, machine learning and automatic speech recognition (ASR) (Rowies & Ghahramani 1999, Frankel 2003). The main purpose of these models is to learn the underlying, typically low dimension, state space to model seemingly complex behaviour in the observation space (data), thus making a distinction between the underlying process and the observations chosen to represent it.

A linear dynamical model (LDM) is described using the following equations (Frankel 2003):

$$\mathbf{y}_t = H\mathbf{x}_t + \boldsymbol{\epsilon}_t \qquad \boldsymbol{\epsilon}_t \sim N(\mathbf{v}, C) \tag{5.1}$$

$$\mathbf{x}_t = F\mathbf{x}_{t-1} + \boldsymbol{\eta}_t \qquad \boldsymbol{\eta}_t \sim N(\mathbf{w}, D) \tag{5.2}$$

where $\mathbf{y}_t$ is an observed feature vector, $\mathbf{x}_t$ is an unobserved (hidden) state vector with initial value at $t = 0$ of $\mathbf{x}_0 \sim N(\pi, \Lambda)$, $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\eta}_t$ are uncorrelated normally distributed noise vectors with means $\mathbf{v}, \mathbf{w}$ and covariance matrices C, D respectively. In Equation 5.1, observation vectors $\mathbf{y_t}$ are a linear transformation of the current state vectors $\mathbf{x_t}$, given by matrix H, corrupted by observation noise $\boldsymbol{\epsilon}_t$, which is Gaussian. The state vectors evolve from one time frame to another according to rotations and stretches via matrix F and also corrupted by the Gaussian state or process noise $\boldsymbol{\eta}_t$, represented by Equation 5.2.

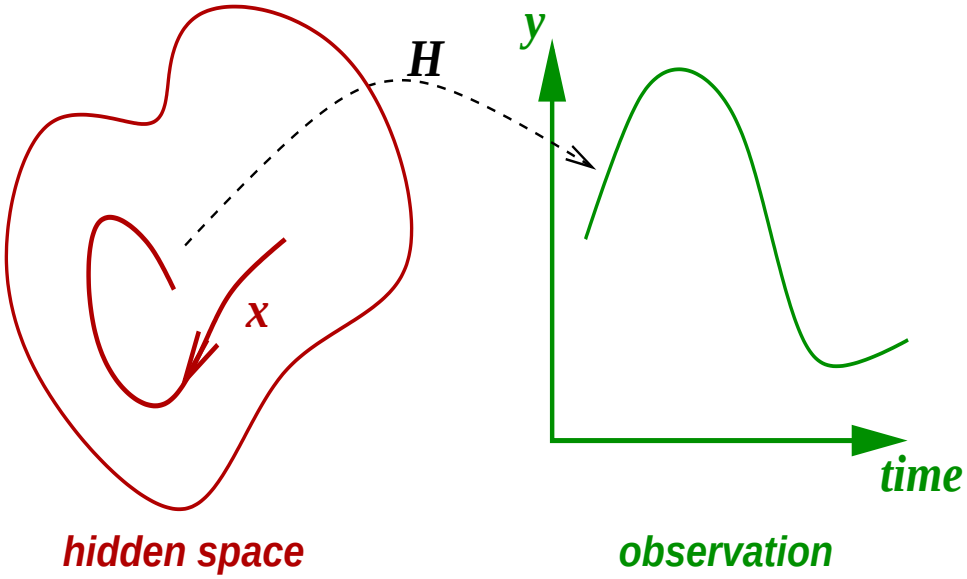

Figure 5.1: In state-space models, the observations are seen as realisations of some unseen, usually lower-dimensional, process (reproduced with the permission from Simon King).

5.2.1 Advantages of using a state space

As mentioned before, the LDM operates in the state space, where it makes smooth and continuous trajectories, as depicted in figure 5.1. Although the state can be considered as the underlying cause of the process, this is hidden and only the output observations are seen. The state can take variety of forms, such as static distributions, long-span auto regressive processes or sets of discrete models (Frankel 2003).

The complexity of the dynamics which the system can model is determined by the dimension of the state vector. For example, a one dimensional state would allow exponential decay or growth with an overall drift, as shown in figure 5.2. This figure shows plots of the state means for two models, one of which exhibits exponential decay and the other exponential growth. Single standard deviations from the mean are also plotted in the figure in dashed lines to demonstrate the change of variance over the time. The parameters used to compute the trajectories were:

- Model a: $F = [0.9], \mathbf{w} = [\mathbf{0.01}], D = [0.01], \boldsymbol{\pi} = [0.7]$ and $\Lambda = [0.01]$

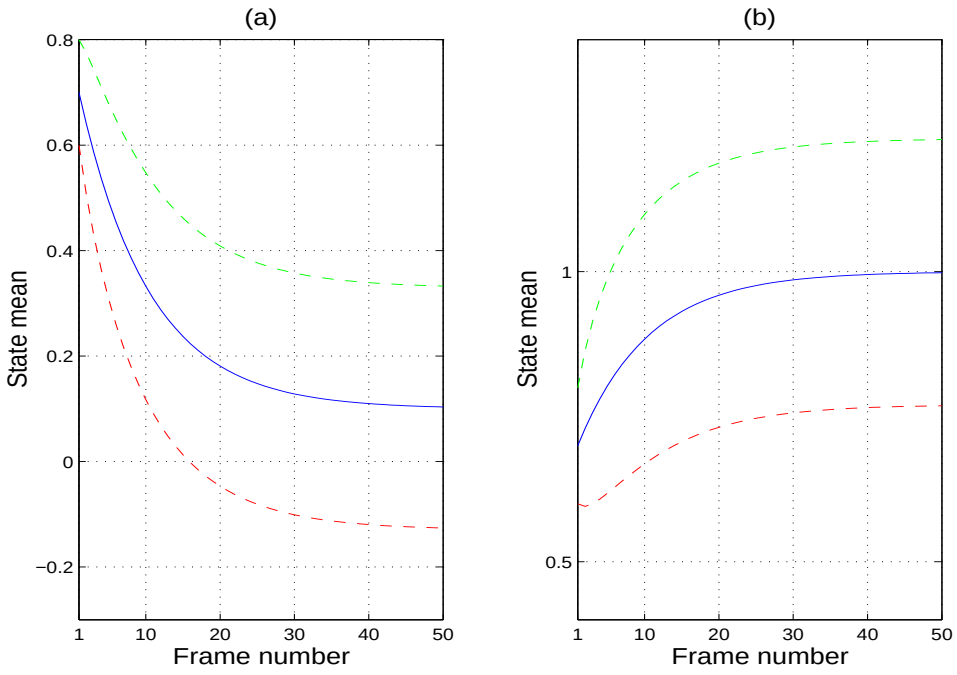

Figure 5.2: Example plots of one dimensional state means showing (a) exponential decay, (b) growth. The dashed lines are placed a single standard deviation from the mean

- Model b: $F = [0.9], \mathbf{w} = [\mathbf{0.1}], D = [0.01], \boldsymbol{\pi} = [0.7]$ and $\Lambda = [0.01]$

In the two-dimensional case, the state process could describe damped or exponentially increasing oscillations again with an overall drift. An example model with a two dimensional state is given to illustrate the state behaviour. Figure 5.3 shows the means of the two state axes over time (frame numbers in the figure) for this model, computed using the following parameters:

$$F = \begin{bmatrix} 0.9 & 0.2 \\ -0.2 & 0.9 \end{bmatrix}, \mathbf{w} = \begin{bmatrix} 0.01 \\ 0.005 \end{bmatrix}, D = \begin{bmatrix} 0.005 & 0 \\ 0 & 0.005 \end{bmatrix},$$

$$\boldsymbol{\pi} = \begin{bmatrix} 0.01 \\ 0.01 \end{bmatrix}, \Lambda = \begin{bmatrix} 0.005 & -0.0004 \\ -0.0004 & 0.005 \end{bmatrix}$$

In figure 5.3, the second plot resembles a phase-shifted re-scaling of the first one. We can see a smooth continuous motion in third plot, where one dimension is plotted

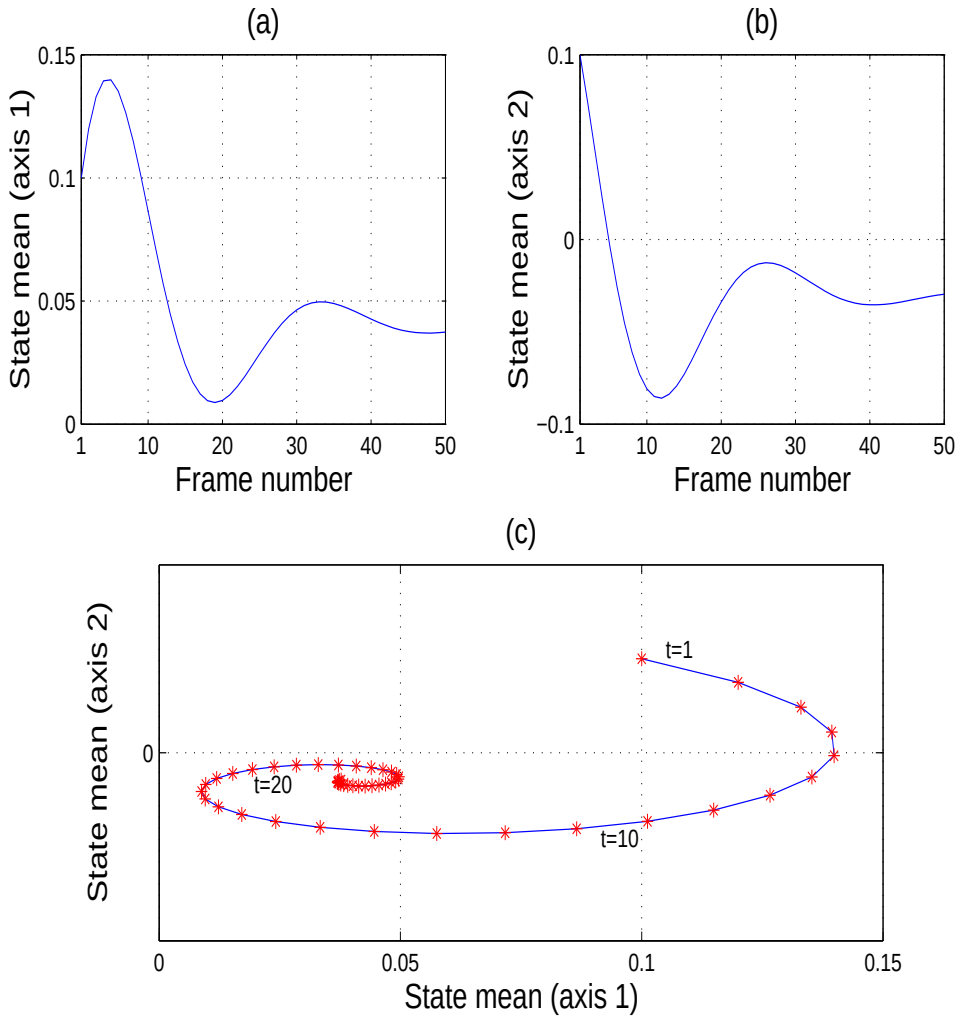

Figure 5.3: Example plots of two dimensional state means showing damped oscillations. (a) state mean (axis 1), (b) state mean (axis 2), (c) Two state means one against another. (adapted from Frankel (2003))

against the other. Higher dimensions allow fairly complex trajectories to be modelled (Frankel 2003).

The above examples demonstrate the advantage of using the underlying state space. This allows any suitable dimension to model the complex behaviour of observation trajectories. However, if we model directly in the observation space it fixes the dimensionality and thus the complexity of the motion. In addition, there may be some redundancy in observation space, which can be eliminated by using a suitable state dimension. The state process is continuous within the phone segment and can also be over boundaries, thus making it appropriate for modelling speech.

5.2.2 Training LDMs

Using LDMs consists of two main tasks: inference and parameter learning. Inference is the task of estimating the hidden state sequence given the observations and parameters. Parameter learning is the task of learning the model parameters given the observations and state sequence.

5.2.2.1 Inference – Kalman filter and RTS smoother

Inference in LDMs involves computing the posterior distribution of the hidden state given an observation sequence ($\mathbf{y}_1^N = \mathbf{y}_1, \cdots, \mathbf{y}_N$) and a set of model parameters. This task consists of two parts:

1. A set of forward recursions known as the **Kalman filter** (Kalman 1960), shown in table 5.1.

2. A set of backward recursions known as Rauch-Tung-Striebel (RTS) smoother (Rauch, Tung & Striebel 1965), shown in table 5.2.

The forward recursions require the initial distribution of the state variables to proceed. Since the state variables are assumed Gaussian, both mean ($\boldsymbol{\pi}$) and covariance (Λ) of initial state are to be specified. These are often set as a zero vector and an identity matrix. The filter first computes the prior state means ($\hat{\mathbf{x}}_{t|t-1}$) and covariances ($\Sigma_{t|t-1}$) (except for the initial state where these are set to $\boldsymbol{\pi}$ and Λ). Then, these estimates are

$$
\begin{aligned}
\hat{\mathbf{x}}_{t|t-1} &= F\hat{\mathbf{x}}_{t-1|t-1} + \mathbf{w} \\
\Sigma_{t|t-1} &= F\Sigma_{t-1|t-1}F^T + D \\
\mathbf{e}_t &= \mathbf{y}_t - \hat{\mathbf{y}}_t = \mathbf{y}_t - \mathbf{v} - H\hat{\mathbf{x}}_{t|t-1} \\
\Sigma_{\mathbf{e}_t} &= H\Sigma_{t|t-1}H^T + C \\
K_t &= \Sigma_{t|t-1}H^T\Sigma_{\mathbf{e_t}}^{-1} \\
\hat{\mathbf{x}}_{t|t} &= \hat{\mathbf{x}}_{t|t-1} + K_t\mathbf{e_t} \\
\Sigma_{t|t} &= \Sigma_{t|t-1} - K_t\Sigma_{\mathbf{e_t}}K_t^T
\end{aligned}
$$

Table 5.1: Kalman filter (forward) recursions

$$
\begin{aligned}
\hat{\mathbf{x}}_{t-1|N} &= \hat{\mathbf{x}}_{t-1|t-1} + A_t(\hat{\mathbf{x}}_{t|N} - \hat{\mathbf{x}}_{t|t-1}) \\
\Sigma_{t-1|N} &= A_t(\Sigma_{t-1|N} - \Sigma_{t|t-1})A_t^T \\
A_t &= \Sigma_{t-1|t-1}F^T\Sigma_{t|t-1}^{-1}
\end{aligned}
$$

Table 5.2: RTS smoother (backward) recursions

projected into the observation space and the error ($\mathbf{e}_t$) between predicted ($\hat{\mathbf{y}}_t$) and actual ($\mathbf{y}_t$) observations is calculated. The Kalman gain (K_t) computed from this error is used to correct the predictions. Finally, the state posteriors ($\hat{\mathbf{x}}_{t|t}$ and $\Sigma_{t|t}$) are estimated.

During the backward recursions (RTS smoother), the state mean ($\hat{\mathbf{x}}_{t|N}$) and covariance matrix ($\Sigma_{t|N}$) are estimated by iterating backwards from the end to the beginning of the N-length observation sequence, i.e. giving estimates which account for past, present and future observations. The matrix A_t provides the weighting for the contribution of the filter.

5.2.2.2 Parameter learning using EM algorithm

The well-known Expectation Maximisation (EM) algorithm can be used for parameter learning. The EM algorithm was introduced in Dempster, Laird & Rubin (1977), for maximum likelihood estimation of incomplete data.

In the case of LDMs, the state is unobserved and it can be treated as missing data. Hence, the standard EM algorithm is an appropriate approach for maximum likelihood estimation. Shumway & Stoffer (Shumway & Stoffer 1982) first presented the

EM algorithm for linear dynamical systems, where the observation matrix (H) is known. Later, several researchers extended this work and presented the EM algorithm with H unknown (Digilakis et al. 1993, Ghahramani & Hinton 1996).

Here, we have used the EM algorithm to compute maximum-likelihood estimates for the model parameters $\{H, F, \mathbf{v}, \mathbf{w}, C, D, \mathbf{x}_0\}$. During the E step, statistics are accumulated over these training examples using the previous set of model parameters. Then in the M-step, these statistics are used to update the model parameters. Refer to Digilakis et al. (1993) and Frankel (2003) for full details.

5.2.2.3 EM initialisation schemes

Though the EM algorithm guarantees to increase the likelihood in each iteration, it may not learn optimal parameters, partly due to its sensitive nature to initial conditions. Hence, inappropriate initial conditions may cause EM to converge towards local rather than the global maxima and impact on final model performance. Recently, various initialisation schemes have been compared in the context of speech modelling (Smith et al. 1999) and phone classification tasks in speech recognition (Frankel 2003). However, there is no standard initialisation method for EM to use in LDM training.

In this thesis, prior to EM, the following three different schemes for initialising model parameters were experimented with:

AR(1) A first order autoregressive (AR) process with some randomness introduced into the estimation was used to initialise the model parameters. This was presented in the software in Ghahramani & Hinton (1996) and it is best used when the dimensions of the state space and observations are the same (Smith et al. 1999). Here, we used a modified version of this software with state dimension equal to observation dimension.

Factor analysis (FA) In this method, a factor analysis model was used to initialise the observation process parameters ($H, \mathbf{v}, C$). Once we initialise v to be the data mean, EM can be used to estimate H and C. Then, these parameters are used to infer the state-space equation parameters ($F, \mathbf{w}, D$). For corresponding derivations and formulae refer to Rowies & Ghahramani (1999) and Frankel (2003).

Empirical Hand-picked and tuned initial values for the model parameters were chosen based on the experiments by Frankel (2003). Some assumptions were made about the function of each parameter and various values were tested in phone classification. It has been observed (Frankel 2003) that values chosen this way performed better than the above factor analyser method.

5.3 Objective measure – join cost

The models were trained on natural speech: they learn the dynamical properties of LSFs from natural speech segments of a particular phone. They typically needed 3-4 iterations for EM to converge. Models were phone specific, with one set of parameters $H, F, C, D, \mathbf{v}, \mathbf{w}$ and $\mathbf{x}_o$ per phone. This is convenient, since joins are usually at the centres of phones in concatenative synthesis, so we can "run" a model from the start of a phone, through the join, to the end of that phone.

Now, to understand how this helps us compute the closeness between the joined phone and natural speech, it is worth thinking of the LDM as tracking the observation (LSF) trajectories. At the start of the phone, the model will follow these trajectories closely since the speech is natural, i.e., it is very much like the data used in training the model. And around the join, the model treats any discontinuity in the LSF trajectories as noise and infers a most likely path through the join. Towards the end of the phone, the model once again follows the observations closely since the speech is natural. The error between the predicted and the actual observations forms the basis of our join cost function.

To find this error, first we compute the log likelihood of the observation sequence $\mathbf{Y}$, given the parameters of model m as follows (Digilakis et al. 1993):

$$log\ p(\mathbf{Y}|m) = -\sum_{t=t_{start}}^{t=t_{end}} \{log|\Sigma_{\mathbf{e}_t}| + \mathbf{e}_t^T \Sigma_{\mathbf{e}_t}^{-1} \mathbf{e}_t\} + const. \tag{5.3}$$

where $\mathbf{e}_t$ and $\Sigma_{\mathbf{e}_t}$ are the prediction error and its covariance for model m, and can be obtained from the standard Kalman filter recursions.

Here, two examples, one with a natural (good) join and other with an artificial (bad) join, are considered to illustrate the behaviour of model log likelihood parameter (pre-

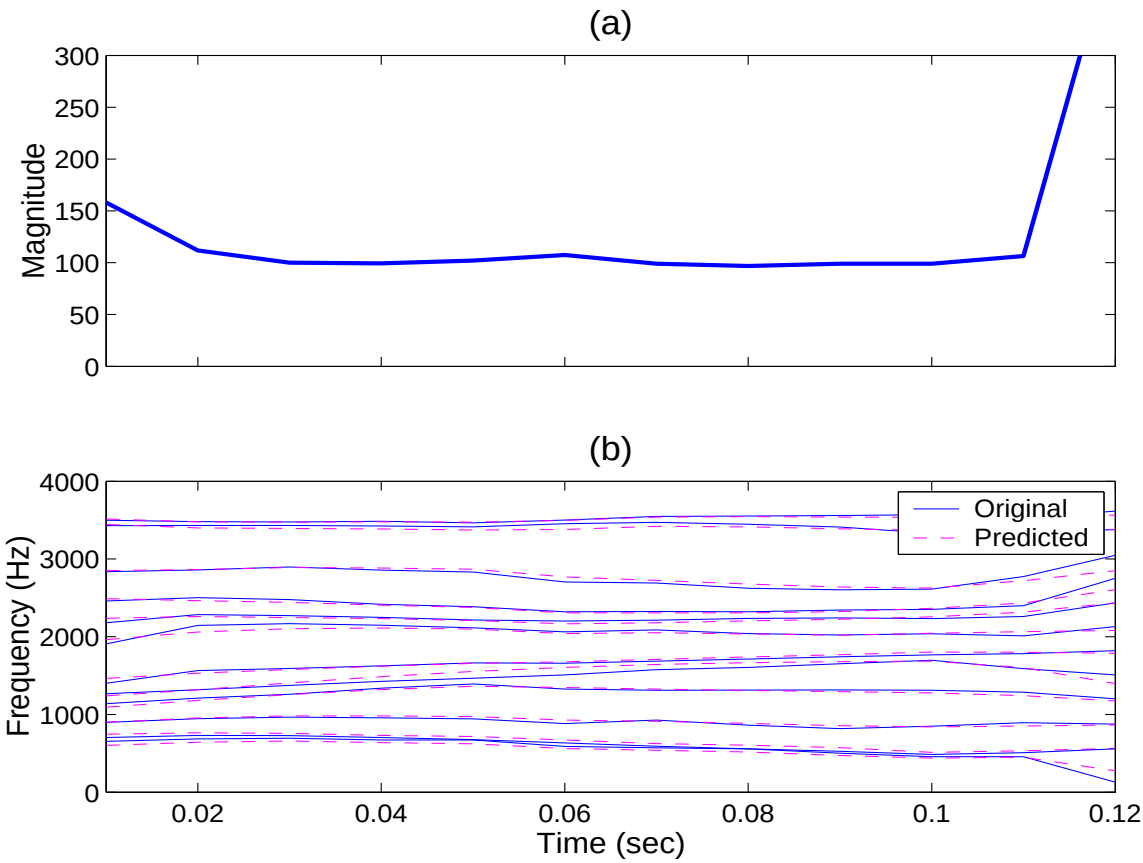

Figure 5.4: (a) Negative log likelihood estimate for a **good** join, (b) Original and predicted observations (LSF)

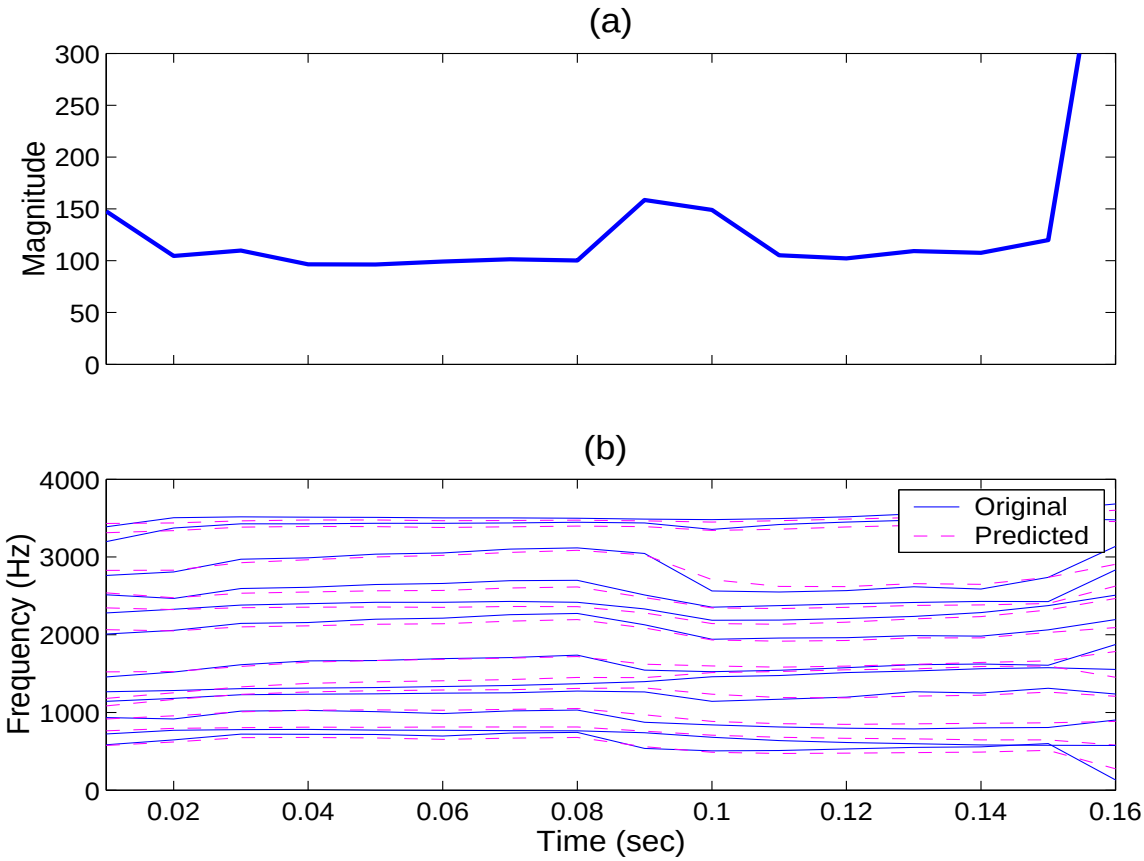

Figure 5.5: (a) Negative log likelihood estimate for a **bad** join, (b) Original and predicted observations (LSF)

diction error). The upper halves of figures 5.4 and 5.5 show the negative log likelihood estimate[2] for a good join and bad join respectively. The increased model prediction error can be clearly seen in figure 5.5 where, in the region of the join, the model infers an optimal trajectory through the LSF parameters (lower half of the figure) but accumulates error between this trajectory and the actual observations. The high prediction errors at the start and end are boundary effects and can be ignored.

5.3.1 Analytical measures

An objective measure to use in the join cost function can be derived from the shape of the negative log likelihood plot, which has a peak around the join point. We have tried three different methods which are geometrically drawn in figure 5.6. These are:

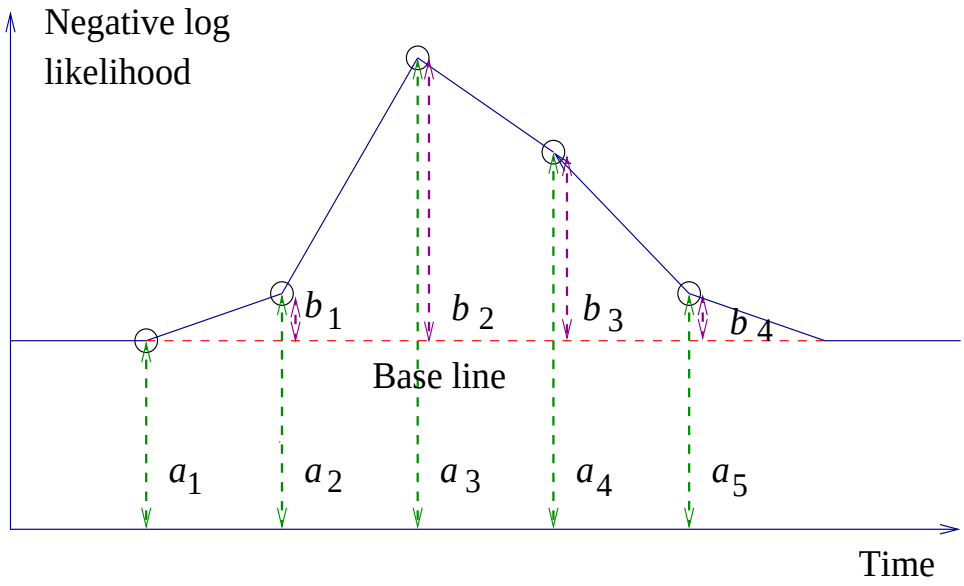

Figure 5.6: A schematic diagram showing three analytical measures

1. An average of the log likelihood over 5 frames centred on the join.

$$\text{Measure1} = \frac{1}{5}\sum_{i=1}^{5} a_i \tag{5.4}$$

2. The relative increase (over an estimated baseline) in the negative log likelihood

[2]For clarity, a high value means more discontinuity

averaged over 3 frames centred on the frame with greatest error;

$$\text{Measure2} = \sum_{i=1}^{3} b_i \tag{5.5}$$

3. This is same as measure 2, except averaged over 4 frames;

$$\text{Measure3} = \sum_{i=1}^{4} b_i \tag{5.6}$$

Measures 2 and 3 are motivated by the plots in the upper halves of figures 5.4 and 5.5 – they measure the area of the "lobe" in figure 5.5.

5.4 Correlation results

Once we have the models trained on line spectral frequencies of natural speech, the analytical measures for joins in all our test stimuli (i.e. around 30 versions for each sentence as discussed in chapter 3) are computed. This is done by running the model on a diphthong at the join of two diphones. Then, correlation coefficients between these analytical measures and mean listener scores were calculated.

The correlation results obtained with three analytical measures using different models (i.e. obtained using three different initialisation schemes prior to EM in training) with various state dimensions are presented in this section.

5.4.1 Effect of state dimension

Choosing the correct dimension for the state space is very important. We trained models with various state dimensions from 1 to 20 to see the effect of state dimension on correlations. Figures 5.7, 5.8, 5.9, 5.10 and 5.11 depict correlation plots for each of the two sentences of five diphthongs used in our perceptual test. The models were initialised using *empirical values* (obtained from experiments by Frankel (2003)) prior to EM and trained on line spectral frequencies (LSF) obtained from original speech with both 8kHz and 16kHz sampling frequencies. In each plot the three analytical measures discussed in the previous section are compared. Also, the highest correlation values are displayed in the legend.

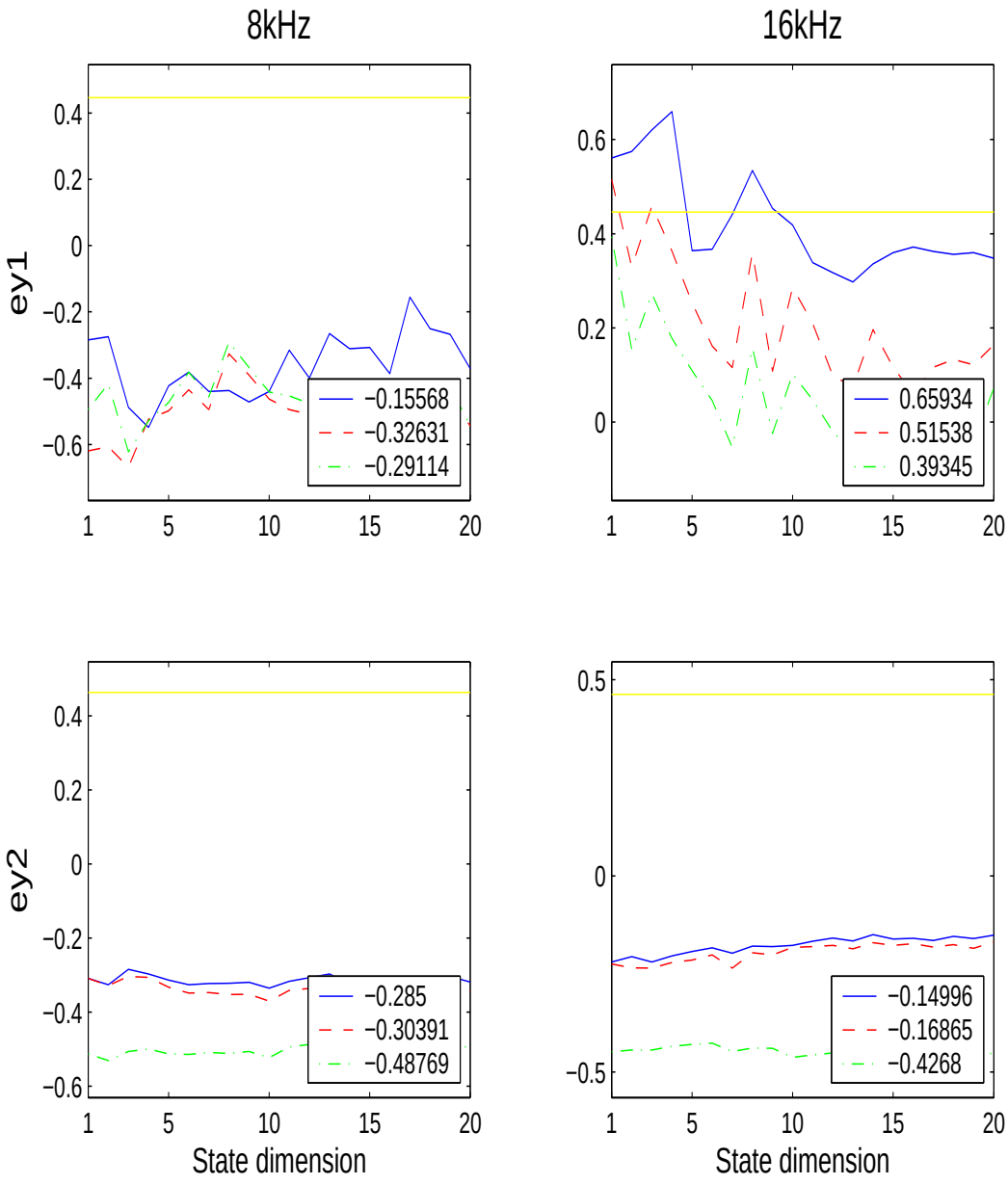

Figure 5.7: Correlation coefficients obtained from models with different state dimensions for diphthong **ey** and for three analytical measures (measure1 (solid blue), measure2 (dashed red) and measure3 (dashdot green)). The yellow line shows the absolute value of correlation below which the correlation is insignificant. Highest values in each case are displayed in the legend.

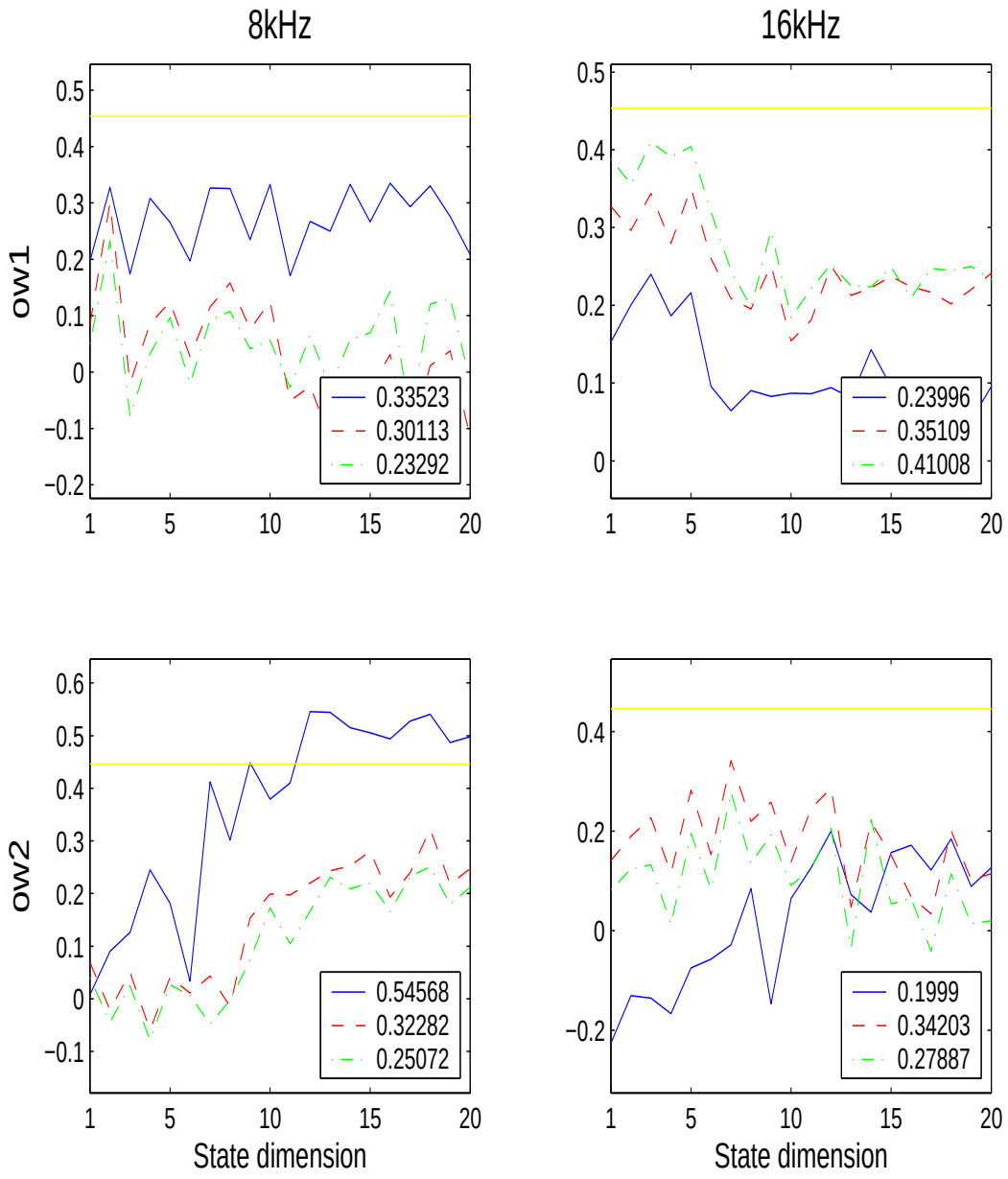

Figure 5.8: Correlation coefficients obtained from models with different state dimensions for diphthong **ow** and for three analytical measures (measure1 (solid blue), measure2 (dashed red) and measure3 (dashdot green)). The yellow line shows the absolute value of correlation below which the correlation is insignificant. Highest values in each case are displayed in the legend.

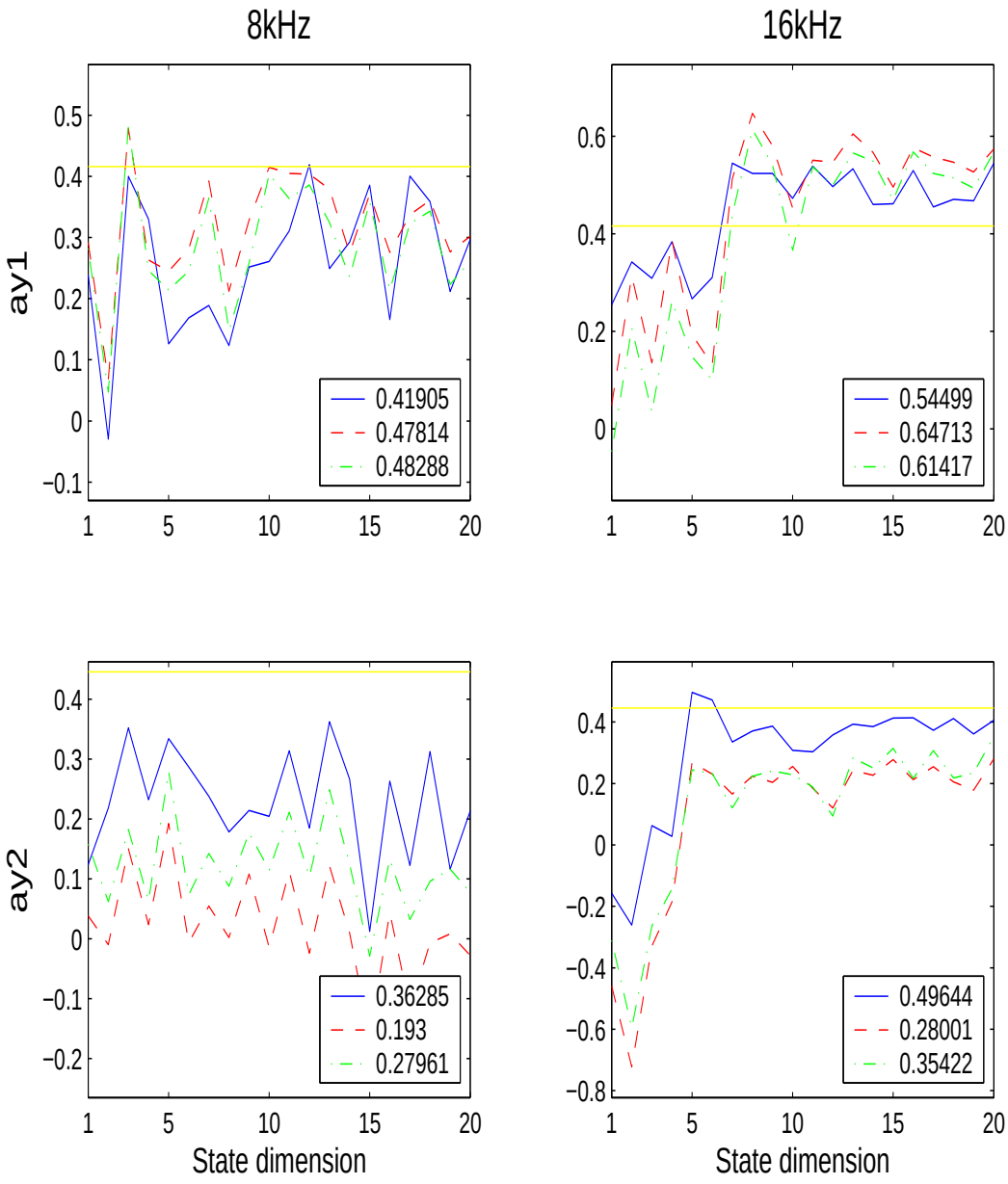

Figure 5.9: Correlation coefficients obtained from models with different state dimensions for diphthong **ay** and for three analytical measures (measure1 (solid blue), measure2 (dashed red) and measure3 (dashdot green)). The yellow line shows the absolute value of correlation below which the correlation is insignificant. Highest values in each case are displayed in the legend.

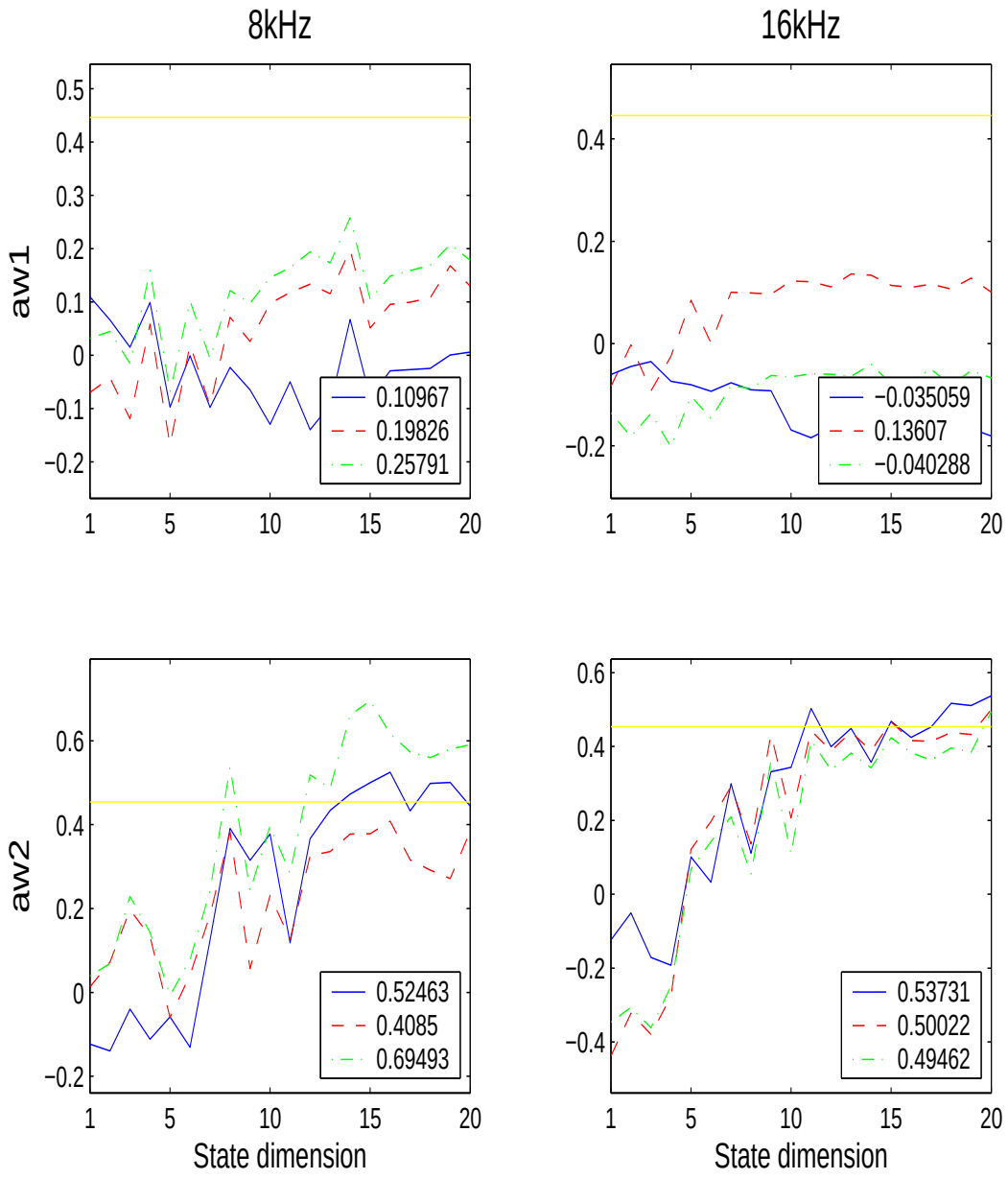

Figure 5.10: Correlation coefficients obtained from models with different state dimensions for diphthong **aw** and for three analytical measures (measure1 (solid blue), measure2 (dashed red) and measure3 (dashdot green)). The yellow line shows the absolute value of correlation below which the correlation is insignificant. Highest values in each case are displayed in the legend.

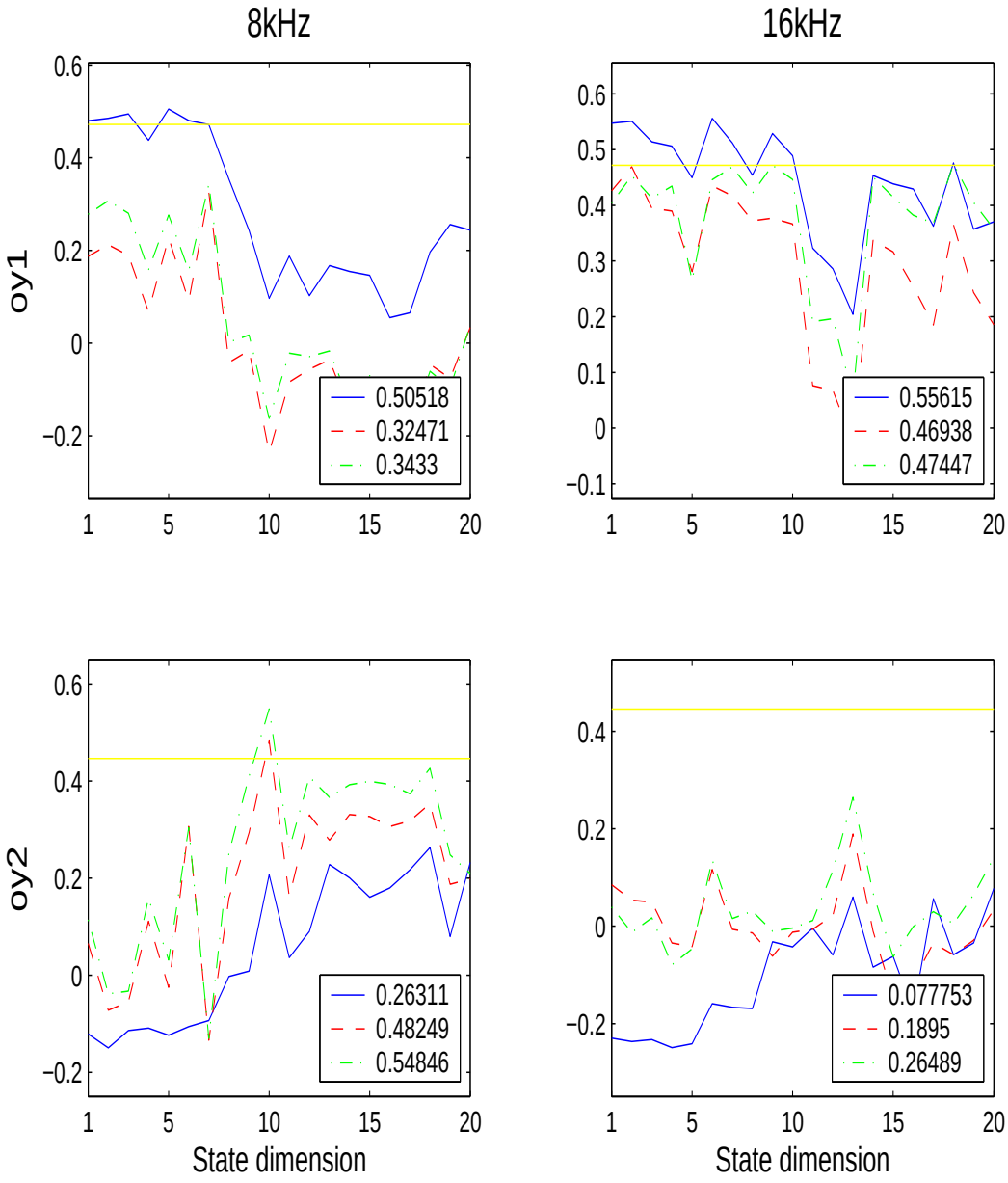

Figure 5.11: Correlation coefficients obtained from models with different state dimensions for diphthong **oy** and for three analytical measures (measure1 (solid blue), measure2 (dashed red) and measure3 (dashdot green)). The yellow line shows the absolute value of correlation below which the correlation is insignificant. Highest values in each case are displayed in the legend.

In general, the correlation plots do not show any consistent trend, for example, in the *ow2* case (for both 8kHz and 16kHz), correlation values are increasing with respect to state dimension and for *ey1* case (16kHz), these are decreasing as state dimension increases. In other-words, a low state dimension yields high correlations in some cases; in other cases, a higher state dimension is required. The first analytical measure (solid blue line in plots) performs better than the other two measures. LDMs trained with LSFs computed on 16kHz sampled speech yielded good correlations compared their 8kHz counterparts. Surprisingly, in *ey1* case (figure 5.7), the 8kHz LSFs produced negative correlations, whereas 16kHz LSFs resulted in positive correlations.

5.4.2 Effect of LDM initialisation

As mentioned in section 5.2.2.3, three different initialisation schemes were used to initialise models prior to EM. These were *AR(1): first order auto-regressive process*, *Factor analysis* and *Empirical values*.

	measure1		measure2		measure3	
	8kHz	16kHz	8kHz	16kHz	8kHz	16kHz
ey	-0.39	0.11	-0.43	0.02	-0.41	-0.04
	-0.32	-0.16	-0.37	-0.19	-0.51	-0.47
ow	0.33	0.10	0.22	0.22	0.16	0.26
	0.44	0.11	0.11	0.24	0.09	0.17
ay	0.23	**0.49**	0.34	**0.49**	0.34	**0.46**
	0.24	**0.54**	-0.04	0.36	0.11	0.41
aw	-0.08	-0.12	-0.02	0.14	0.05	-0.07
	0.49	0.41	0.25	0.41	**0.54**	0.36
oy	0.07	0.21	-0.15	0.04	-0.13	0.19
	0.04	-0.07	0.05	-0.06	0.10	0.03

Table 5.3: Correlations between perceptual scores and three measures based on a LDM estimate, using the AR(1) method to initialise the model parameters prior to EM.

Tables 5.3, 5.4 and 5.5 report the correlation coefficients of the three types of our

analytical measures using likelihood estimates with mean listener preference ratings. The values 8kHz and 16kHz are referred to models trained on LSFs obtained from 8kHz and 16kHz sampled speech. The correlation value changes with the state dimension as discussed in above section, hence only the best result for each case is quoted in the tables. Correlation coefficients above the 1% significance level are in bold.

	measure1		measure2		measure3	
	8kHz	16kHz	8kHz	16kHz	8kHz	16kHz
ey	-0.26	**0.61**	-0.27	**0.57**	-0.28	0.35
	-0.30	-0.14	-0.32	-0.16	-0.50	-0.43
ow	0.35	0.20	0.11	0.42	0.11	**0.46**
	0.48	0.30	0.27	0.28	0.28	0.19
ay	0.40	**0.56**	0.39	**0.54**	0.41	**0.51**
	0.36	**0.52**	0.13	0.34	0.33	0.43
aw	0.10	-0.03	0.01	0.13	0.09	-0.04
	0.58	**0.55**	0.30	**0.49**	**0.68**	**0.48**
oy	**0.48**	**0.55**	0.26	0.43	0.35	0.44
	0.03	-0.07	0.07	0.10	0.09	0.06

Table 5.4: Correlations between perceptual scores and three measures based on a LDM estimate, using a factor analyser to initialise the model parameters prior to EM.

Comparison of tables 5.3, 5.4 and 5.5 shows that there is no significant difference in absolute correlation values obtained using models with different initialisations. However, the number of 1% significant correlations were more for models initialised using a factor analyser or our hand-picked values than using an AR(1) model. Also, it is clear that measure1 is the best among our three analytical measures in terms of the number of 1% significant correlations. This measure uses a simple average of the absolute model error over 5 frames centred on the join. The other measures attempt to calculate the "extra" error – the lobe seen in the upper half of figure 5.5 – which we hypothesised would be a good indicator of the difference between the joined phone and a natural token. This hypothesis is not supported by our results.

	measure1		measure2		measure3	
	8kHz	16kHz	8kHz	16kHz	8kHz	16kHz
ey	-0.15	**0.66**	-0.32	**0.51**	-0.29	0.39
	-0.28	-0.15	-0.30	-0.17	-0.49	-0.43
ow	0.33	0.24	0.30	0.35	0.23	0.41
	0.55	0.20	0.32	0.34	0.25	0.28
ay	**0.42**	**0.54**	**0.48**	**0.65**	**0.48**	**0.61**
	0.36	**0.50**	0.19	0.28	0.28	0.35
aw	0.11	-0.03	0.20	0.14	0.26	-0.04
	0.52	**0.54**	0.41	**0.50**	**0.69**	**0.49**
oy	**0.50**	**0.56**	0.32	**0.47**	0.34	**0.47**
	0.26	0.06	**0.48**	0.19	**0.55**	0.26

Table 5.5: Correlations between perceptual scores and three measures based on a LDM estimate, using hand-picked values to initialise model parameters prior to EM.

5.4.3 Models trained on Bark-scaled LSFs

We also trained LDMs using Bark-scaled LSFs computed using natural speech of 8kHz and 16kHz sampling frequency. The results were not better than those obtained from models trained on the original LSFs.

5.5 Conclusions

A new method for computing join cost has been introduced in this chapter, which uses a linear dynamical model (LDM) to model line spectral frequencies (LSF) obtained from natural speech. This model uses an underlaying hidden state space in which it makes smooth and continuous trajectories. This state space can be seen as an analogy for underlying articulator movements in speech production.

Three different initialisation schemes – *AR(1): first order auto-regressive process*, *Factor analysis* and *Empirical values* – were used prior to Expectation-Maximisation (EM) in training the LDMs. The trained models ran on the joined phones and the

objective join cost function was defined based on log likelihood of the observations given the model. Based on the shape of these likelihood plots, three different analytical measures were derived. Then, correlation values were computed between these measures and mean listener scores. The main observations from these results can be summarised as follows:

- The correlations with perceptual scores do not show any consistent trend in state dimensions
- Initialising the model parameters using a *factor analyser* or our *empirical values* yields better correlations than using an AR(1) model; overall *empirical values* resulted more 1% significant correlations
- The measure1 is best among our three analytical measures
  - *this measure uses a simple average of the log likelihood over 5 frames centred on the join*
- There is no improvement observed using models trained on Bark-scaled LSFs. In fact their performance is poor compared to models trained on original LSFs.
- Models trained using LSFs obtained from 16kHz sampled speech yielded good results compared their 8kHz counterparts.

Chapter 6

Subjective Evaluation of Join Cost & Smoothing

6.1 Introduction

In the previous chapters, we have proposed join cost functions derived from spectral distances, which have good correlations with perceptual scores obtained for a range of concatenation discontinuities. To further validate their ability to predict concatenation discontinuities, we have chosen the best three spectral distances from the previous three chapters. Then, we have compared them subjectively in a listening test using synthetic stimuli, where units are selected using *rVoice* with a different voice, *a female voice*[1].

Various combinations of the three join cost functions and different smoothing methods are compared in the listening test to discover which of these join cost functions and smoothing methods perform better. This test is also intended to discover whether the smoothed line spectral frequencies (LSFs) obtained from the Kalman filter produce better synthesis than the linearly smoothed LSFs. In order to use different smoothing operations, we implemented[2] the residual excited linear prediction (RELP) synthesis method for waveform generation of the units selected from *rVoice* synthesis system.

We start this chapter with a description of the join cost functions and smoothing

[1] In previous chapters a male voice was used

[2] We did not use *rVoice* for waveform generation as we have no access to its source code and can only plug-in join cost code.

methods used. Also, we explain the implementation of the RELP synthesis method used for waveform generation. In section 6.3, the design and procedure of the listening test is discussed. Finally, we present subjective results of these various combinations and discuss them in section 6.4.

6.2 Join cost functions and smoothing techniques

6.2.1 Join cost functions

We have chosen three of the best spectral distances from chapters 3 and 5. These are used as the join cost functions in unit selection search. These three spectral distance measures and our names for the join cost functions derived from them are as follows:

1. *Mahalanobis distance on line spectral frequencies (LSF) and their deltas of frames at the join. The join cost function based on this is termed* ***LSF join cost****.*

2. *Mahalanobis distance computed using multiple centroid analysis (MCA) coefficients of multi-frames (seven frames, i.e. three frames on either side of join plus one frame at the join). The join cost function based on this is termed* ***MCA join cost****.*

3. *The join cost derived from the negative log likelihood estimated by running the Kalman-filter on LSFs of the phone at the join is termed* ***Kalman join cost****.*

The above are used as the join cost function with proper scaling[3]. Since the focus of this thesis is on the spectral component of the join cost function, F0 and energy components are not considered. The first join cost function produced **six** 1% significant correlations (table 3.5). We achieved **seven** 1% significant correlations for the second measure (table 3.17). The Kalman-based join cost produced a maximum of **five** 1% significant correlations (tables 5.4 and 5.5). As mentioned in section 3.3, we used the number of 1% significant correlations to compare these three join costs. The rankings of these three join costs are shown in table 6.1.

[3]The join costs from these functions are very high compared to target costs and if we use them "as is" then join cost and target cost contributions will not be balanced in unit selection and result in poor units being selected

Rank	*Join Cost*
1	MCA join cost
2	LSF join cost
3	Kalman join cost

Table 6.1: Rankings for three join costs, obtained in the first listening test

6.2.2 Smoothing techniques

After units are concatenated, most systems attempt some form of local parameter smoothing to disguise the remaining discontinuity. One of our goals is to combine the join cost process and the join smoothing operation in some optimal way as these two processes interact closely. Suppose, a large database and a perfect join cost function available then no smoothing would be required. On the contrary, the join cost function would be less important if we could smooth joins better.

As mentioned in the previous chapter, the linear dynamic models (LDM) can also smooth the observations (LSFs in our case). Hence, we run the model on LSFs of the phones at the join in both forward and backward directions to compute the log likelihood of observations given the model, and to obtain smoothed LSFs. These smoothed LSFs are then used in RELP synthesis to generate synthetic waveform. We are investigating the combined Kalman filter based join cost function and Kalman smoothing operation as one possible approach towards the above objective. So, in the listening test, we also compare the Kalman smoothing operation with a linear smoothing technique (Dutoit 1997).

6.2.2.1 Linear smoothing

The line spectral frequencies (LSF) have good interpolation properties and yield stable filters after interpolation (Paliwal & Kleijn 1995). Although LSF interpolation is widely used in speech coding, it can also be used for speech synthesis. Dutoit (1997) showed that LSFs have good interpolation properties and produce smoother transitions. LSF interpolation was compared with other smoothing methods by Chappell & Hansen (2002) and they observed good performance in many cases.

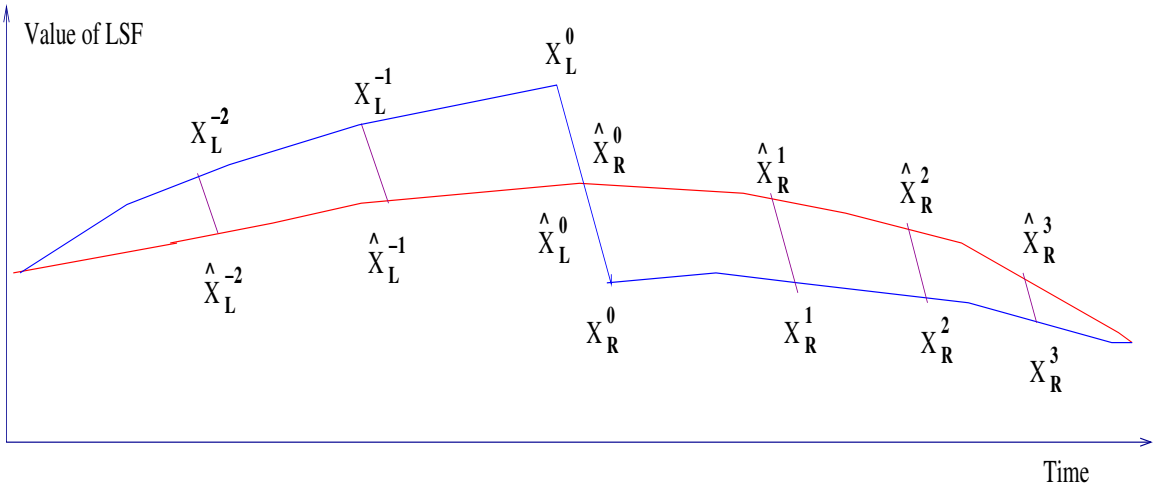

Figure 6.1: Linear smoothing on parameters (LSFs) of frames at the join (adapted from Dutoit (1997)).

We have implemented linear smoothing on LSFs of a few frames of the phones at the join as presented in (Dutoit 1997). The main idea of this technique is to distribute the difference of the LSF vectors at the join across a few frames on either side of the join. To explain this technique, consider L and R as left and right segments at the join and X is a LSF vector $X_1, X_2, \cdots, X_N$. Assume the number of frames on the left side and the right side of the join to be M_L and M_R respectively. Then, the LSFs after smoothing ($\hat{X}$) are:

$$\hat{X}_L^{-i} = X_L^i + (X_R^0 - X_L^0)\,\frac{M_L - i}{2M_L} \qquad 0 \le i < M_L \tag{6.1}$$

$$\hat{X}_R^{j} = X_R^j + (X_L^0 - X_R^0)\,\frac{M_R - j}{2M_L} \qquad 0 \le j < M_R \tag{6.2}$$

where X_L^0 and X_R^0 are frames at the end of L and beginning of R, i.e. exactly at the join. The function of linear smoothing is showed in figure 6.1, where M_L and M_R are 2 and 3 respectively.

6.2.3 Residual excited linear prediction (RELP) based synthesis

Residual excited LPC (RELP) is one of the standard methods for resynthesis, which is also used in Festival (Black & Taylor 1997*b*). In this method, first LPC analysis has to be carried out on the original speech to obtain LPC parameters. Then, inverse filtering

is performed to get the residual signal. Consider original speech sample $x[n]$ which can be predicted as a linear combination of the previous p (linear prediction order) samples, as given below:

$$\tilde{x}[n] = \sum_{i=1}^{p} -a_i \; x[n-i] \tag{6.3}$$

where a_i are prediction coefficients and $x[n-i]$ are past speech samples. The prediction error due to this approximation is:

$$e[n] = x[n] - \tilde{x}[n] = x[n] + \sum_{i=1}^{p} a_i \; x[n-i] \tag{6.4}$$

This error is known as the *residual signal*, which can be used as the excitation to the LPC filter to get a near-perfect reconstruction of the speech signal.

During LPC analysis we have computed the LPC parameters using asymmetric[4] hanning-windowed pitch-synchronous frames of the original speech as shown in figure 6.2. The advantage of using the asymmetric window can be observed in the figure, where successive pitch periods are very different in size and the window is not centered. The sample plots shown in the figure are two pitch periods in length. The residual is computed by passing the windowed original speech (plot (c)) through the inverse LPC filter. A sample residual signal is depicted in plot (d) of the figure 6.2.

Once the units are selected using the *rVoice* synthesis system, the corresponding LPCs and residual signals from the database are assembled. We convert the LPC parameters to LSFs, then employ smoothing methods (linear and Kalman) and then convert back to LPC parameters for synthesis. The residual is not modified by the smoothing operation. Then, the LPC filter is excited using the residual to reconstruct the output speech waveform. In figure 6.2, the output waveform is depicted in the last plot, which is a near-perfect reconstruction of the original signal. To get the full synthetic waveform for an utterance we overlap and add these two-pitch-period output waveforms.

[4]The left and right halves of the window are different.

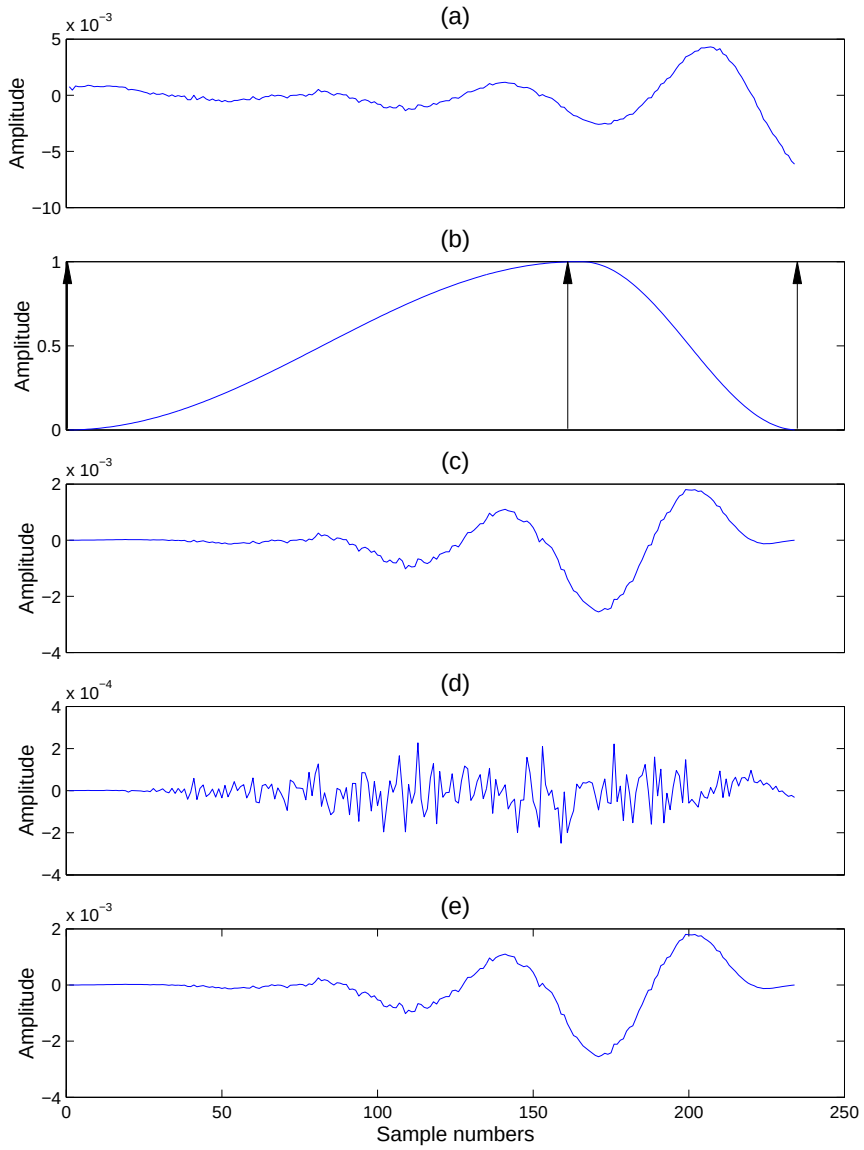

Figure 6.2: RELP synthesis using an asymmetric window: (a) Original waveform (b) Asymmetric hanning window (pitch marks shown as arrows) (c) Windowed original waveform (d) Residual signal (e) Reconstructed waveform

6.3 Listening test

A listening test was designed to evaluate the three join costs and the above smoothing methods, and to compare the smoothed LSFs obtained from Kalman filter and linear smoothing on LSFs. We are testing the following three things:

- *Compare three join costs; LSF join cost, MCA join cost and Kalman join cost; irrespective of smoothing methods*
- *Similarly, compare three smoothing methods; no smoothing, linear smoothing and Kalman smoothing; irrespective of join costs*
- *Check if Kalman join cost together with Kalman smoothing is any better than LSF join cost with linear smoothing*

6.3.1 Test design & stimuli

To describe our test design, we use 1,2 and 3 to denote the three join costs: LSF, MCA and Kalman respectively. The three smoothing methods: a,b and c are no smoothing, linear smoothing and Kalman smoothing in that order. Now, we have 9 different synthetic versions for each of our test sentences obtained with the three join costs and the three smoothing methods, for example V_{1a} means synthesised version using join cost function "1" and smoothing method "a".

Ideally, to know which combination of join cost and smoothing method is the best, we need to compare all the combinations from 9 different versions. Such combinations formed from 9 versions result in 36 pairs[5], as shown in table 6.2, which are divided into 12 symmetric[6] blocks.

To know which join cost performs better, the three blocks in the first row need to be considered. Similarly, to compare smoothing methods three blocks in the second row have to be taken. The remaining two rows (in addition to first and second rows) are required to know which particular join cost and smoothing pair performs better than

[5]Each pair means one comparison, for example $V_{1a} - V_{2a}$

[6]Each block has an equal number of a particular version, for example in the first block V_{1a} appears twice, similarly V_{2a} and V_{3a} appear twice.

V_{1a}-V_{2a}	V_{1b}-V_{2b}	V_{1c}-V_{2c}
V_{2a}-V_{3a}	V_{2b}-V_{3b}	V_{2c}-V_{3c}
V_{3a}-V_{1a}	V_{3b}-V_{1b}	V_{3c}-V_{1c}
V_{1a}-V_{1b}	V_{2a}-V_{2b}	V_{3a}-V_{3b}
V_{1b}-V_{1c}	V_{2b}-V_{2c}	V_{3b}-V_{3c}
V_{1c}-V_{1a}	V_{2c}-V_{2a}	V_{3c}-V_{3a}
V_{1a}-V_{2b}	V_{2a}-V_{3b}	V_{3a}-V_{1b}
V_{2b}-V_{3c}	V_{3b}-V_{1c}	V_{1b}-V_{2c}
V_{3c}-V_{1a}	V_{1c}-V_{2a}	V_{2c}-V_{3a}
V_{1a}-V_{2c}	V_{2a}-V_{3c}	V_{3a}-V_{1c}
V_{2c}-V_{3b}	V_{3c}-V_{1b}	V_{1c}-V_{2b}
V_{3b}-V_{1a}	V_{1b}-V_{2a}	V_{2b}-V_{3a}

Table 6.2: All possible pairwise comparisons

any other possible pair. However, this increases the number of our test stimuli and it is then not possible to test on many sentences.

In other words, if we consider all 36 pairs, a maximum of four sentences can be tested assuming the test duration is 30-40 minutes. In addition, subjects may lose interest after listening to the same sentences many times. To avoid the latter problem, we can rotate the various blocks between different subjects, i.e. presenting only a few (say 3 out of 12) blocks of each sentence and thus increase the number of sentences to each subject. But in this case, we will not get many subjective results per sentence as 4 subjects are used to test one sentence.

Hence we compared only one pair in the last two rows: Kalman join cost and Kalman smoothing *vs* LSF join cost and linear smoothing (i.e. V_{3c} vs V_{1b}). We have chosen linear smoothing since it is a popular and standard procedure in current synthesis systems and we feel combining this with one of our best join costs, the *LSF join cost*, becomes a strong contestant to the V_{3c}. To do this comparison we added the V_{3c} and V_{1b} pair in our test stimuli to the first two rows of table 6.2.

The test sentences used in our listening test are presented in table 6.3. These eight

sentences were selected randomly from twenty such sentences.

Sentence 1	Paragraphs can contain many different kinds of information.
Sentence 2	The aim of argument, or of discussion, should not be victory, but progress.
Sentence 3	He asked which path leads back to the lodge.
Sentence 4	The negotiators worked steadily but slowly to gain approval for the contract.
Sentence 5	Linguists study the science of language.
Sentence 6	The market is an economic indicator.
Sentence 7	The lost document was part of the legacy.
Sentence 8	Tornadoes often destroy acres of farm land.

Table 6.3: Listening test sentences

6.3.2 Test procedure

The listening test is divided into two parts to provide a few minutes break for the subjects. Each part consists of 96 pairs of synthetic stimuli (12 synthetic versions of all sentences listed in the table 6.3) covering the pairs in all blocks of the first two rows in the table 6.2, including one pair ($V_{3c} - V_{1b}$) and some validation pairs, i.e. presenting the above pairs in reverse order ($V_{1b} - V_{3c}$).

In each part, the two rows (table 6.2) including a pair ($V_{3c} - V_{1b}$) and 2 validation pairs are presented alternatively to each subject as shown in figure 6.3. In figure 6.3, R1 and R2 each consist of 12 pairs of synthetic stimuli and covered in two parts (PART1 and PART2) for 8 sentences. The pairs for all sentences were randomised within each part of the test and presented to the subjects. For each pair of stimuli they are asked to judge which one is better by keying 1 or 2. This is a forced choice.

There were 33 participants in this listening test. Most of them were people in CSTR or PhD and MSc students in the dept. of Linguistics with some experience of speech synthesis. Around half of them were native speakers of British English. The tests were conducted in sound-proof booths using headphones. After the first part, the subjects were asked to take a rest for a few minutes. On the average, each part took around 15 minutes and about 30-40 minutes for completion of two parts. The informal feedback

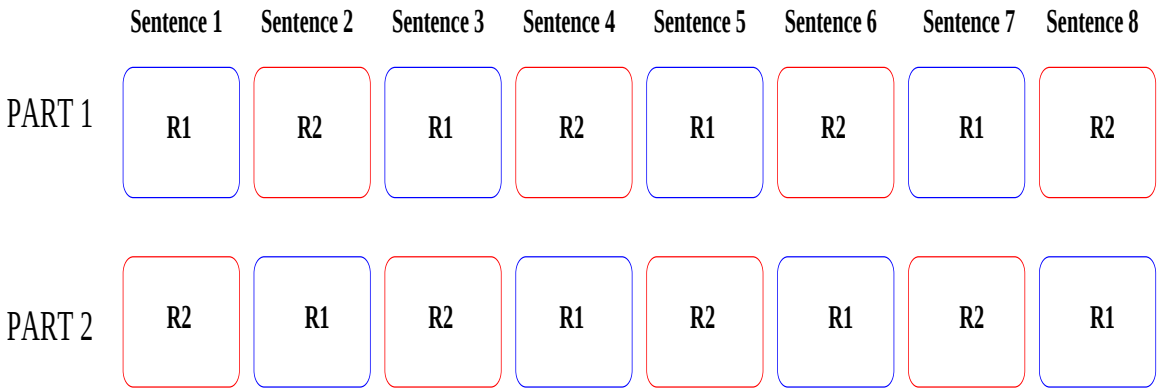

Figure 6.3: Test procedure, in each part the two rows (R1 and R2) are presented alternatively.

from the subjects indicated that there was not much difference between the two stimuli in many pairs. In fact a few of them felt that those pairs were the same, hence found it a difficult task.

6.3.3 Validation procedures

To check the validity of the subjects' results, we included 16 validation pairs in each part of the test. These pairs appear in reverse order. We have adopted a scoring system, where subjects are given a score of 1 or 0 for each of these 16 pairs. If subjects keyed the same response (i.e. 1 or 2) for the original pair and the validation pair then it is an error and they get a score of 0 as they preferred different stimulus in original and validation pairs. If they key opposite responses (for example, 1 for original pair and 2 for validation pair) then they will get a score of 1. These scores are accumulated for 16 pairs for each part of the test. In figure 6.4, we have shown the number of parts which have equal or more validation scores for each validation cutoff ranging from 1 to 16. For example, the number, 37 on top of bar corresponding to the validation cutoff, 10 indicates the number of parts which got a validation scores of 10 or more.

We performed another validation procedure on the block level. Consider the first block in table 6.2; $V_{1a} - V_{2a}$, $V_{2a} - V_{3a}$ and $V_{3a} - V_{1a}$. If subjects preferred all the first stimuli (V_{1a}, V_{2a} and V_{3a}) then the block becomes invalid because, if they prefer V_{1a} and V_{2a}, then for the third pair, the valid selection is V_{1a}. Similarly, they can not prefer all

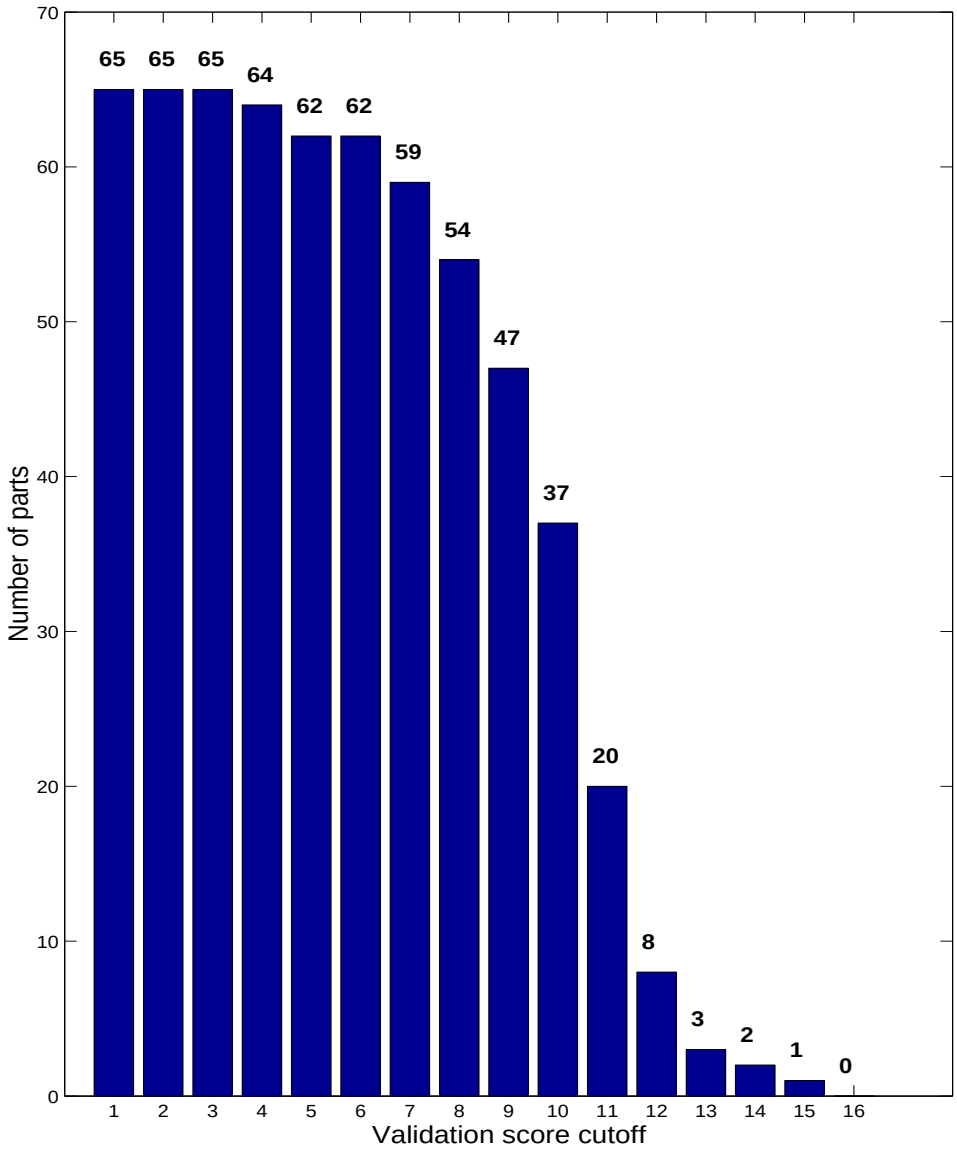

Figure 6.4: Subjects validity

the second stimuli in a block.

6.4 Subjective evaluation

6.4.1 Join costs

In figure 6.5, we show preferences for the three join costs for each sentence using the subjects who got validation scores of 10 or more out of 16. The score 10 is taken as the limit because the chance level is 50%[7] i.e. 8 out of 16. It can be observed from the figure that LSF join cost is preferred more times than MCA join cost and Kalman join cost.

The preferences for the three join costs after removing invalid blocks (as discussed above) are presented in figure 6.6. Though the number of preferences for each sentence is reduced, there is no effect on preference patterns. In other words, the LSF join cost has more preferences compared to MCA join cost and Kalman join cost as observed previously (see figure 6.5).

6.4.1.1 Paired t-test

We conducted a paired t-test to check the significance of these preference ratings. In this test, preferences for join costs for all sentences (each sentence as a group) were considered. The null hypothesis is that the mean difference $\bar{d}$ between the number of preferences is zero; the alternative hypothesis is it is greater than zero ($\bar{d} \neq 0$). The test statistic (t) can be computed as follows (McGhee 1985):

$$t = \frac{\bar{d}}{s/\sqrt{n}} \tag{6.5}$$

where s is the standard error of the differences and n is the number of groups (in our case $n = 8$). The value of t is compared to the critical values of students t-distribution with $n - 1$ degrees of freedom to find the probability by chance or significance level (α)[8]. Low probability values ($\alpha \leq 0.01$) reject the null hypothesis and one can say the

[7] To get valid score (1) the probability is 0.5, as the correct choice is one of the two choices in the validation pair

[8] Significance level is different from p-value, it is the probability that the test statistic being used will have a value at least as extreme as that produced by the sample (McGhee 1985)

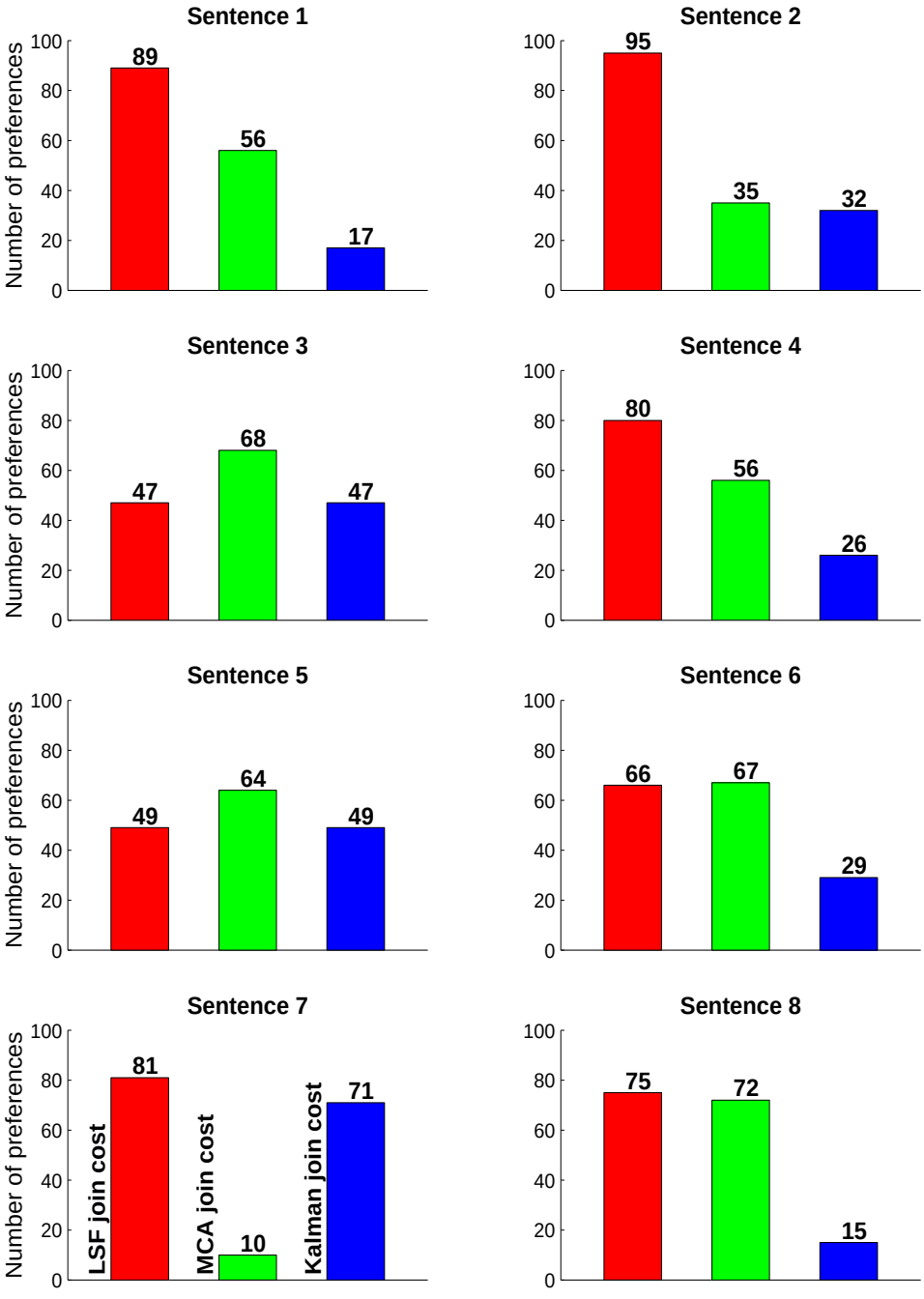

Figure 6.5: Join cost evaluation, validation cutoff is 10

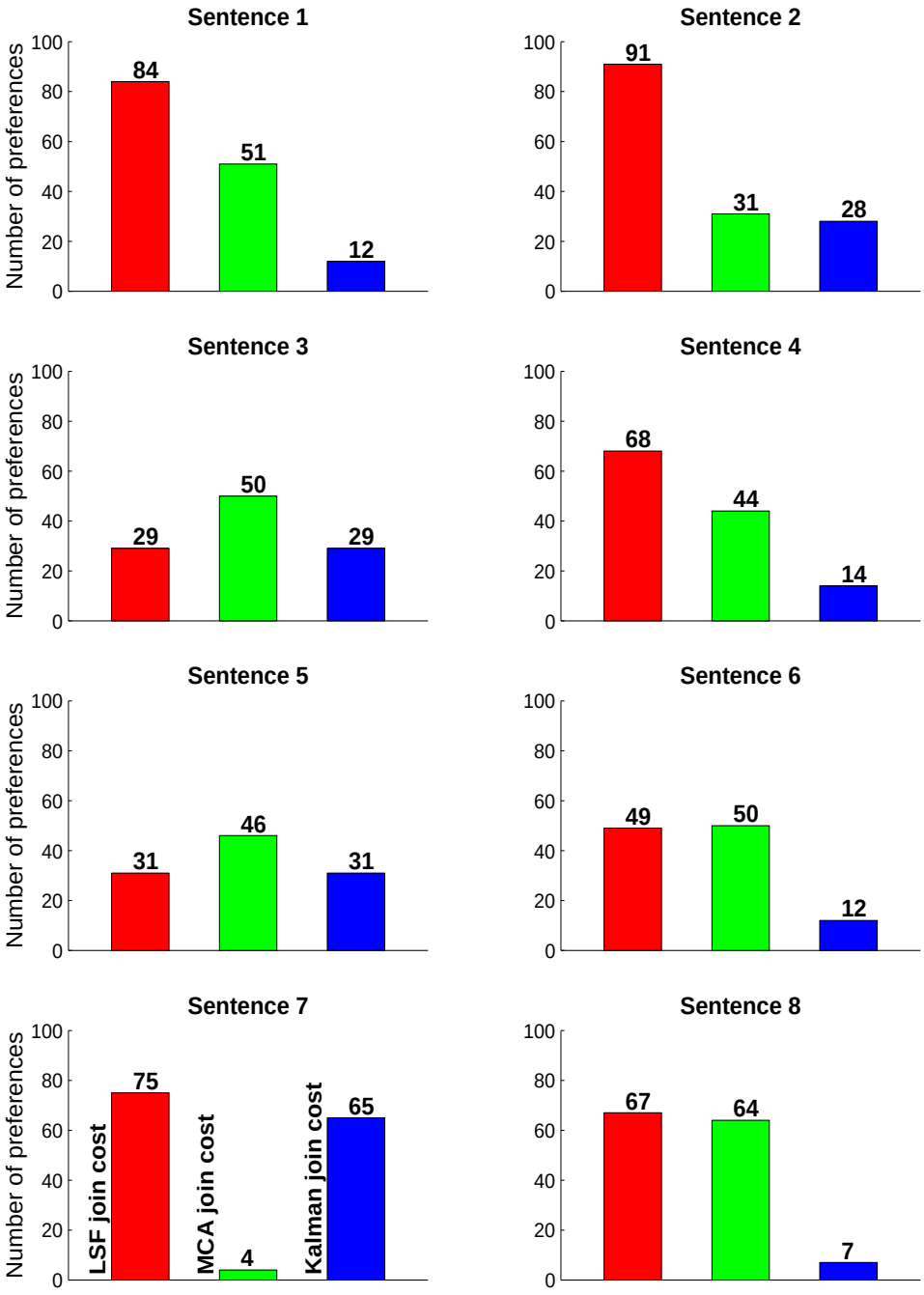

Figure 6.6: Join cost evaluation, validation cutoff is 10 plus block validation check (after removing invalid blocks)

preference to particular join cost is statistically significant.

Validation	*LSF vs MCA*		*MCA vs Kalman*		*Kalman vs LSF*	
cutoff	t	α	t	α	t	α
1	1.6538	0.20	1.5069	0.20	-3.7637	**0.01**
2	1.6538	0.20	1.5069	0.20	-3.7637	**0.01**
3	1.6538	0.20	1.5069	0.20	-3.7637	**0.01**
4	1.6370	0.20	1.5135	0.20	-3.8007	**0.01**
5	1.6816	0.20	1.5193	0.20	-3.9432	**0.01**
6	1.6816	0.20	1.5193	0.20	-3.9432	**0.01**
7	1.6462	0.20	1.5240	0.20	-3.9969	**0.01**
8	1.6631	0.20	1.5510	0.20	-3.8310	**0.01**
9	1.5910	0.20	1.5760	0.20	-3.8367	**0.01**
10	1.6088	0.20	1.4010	> 0.2	-3.5204	**0.01**
11	1.6188	0.20	1.4652	0.20	-3.2729	0.02
12	2.1613	0.10	2.0706	0.10	-3.0818	0.02
13	0.8699	> 0.2	2.2958	0.10	-2.5342	0.05
14	0.7638	> 0.2	2.1573	0.10	-2.4538	0.05
15	0.5401	> 0.2	0.9564	> 0.2	-2.3080	0.10

Table 6.4: Paired t-test statistics for the join costs

A two-tailed t-test was used, since we are looking for a preference on either side. In table 6.4, we present t and α for preference ratings obtained from subjects with validation cutoffs ranging from 1 to 15 (after removing invalid blocks). The preference for LSF join cost over MCA join cost is not statistically significant though the LSF join cost has a greater number of preferences. The preference towards MCA join cost compared to Kalman join cost is also not statistically significant. LSF join cost preferred to Kalman join cost is statistically significant[9] for low validation cutoffs. However, it is less significant for high validation scores (for consistent subject results).

[9]The negative value of t indicates the preference is the other way, i.e. in this case the preference for LSF over Kalman.

6.4.2 Smoothing methods

The preferences for smoothing methods for each sentence are shown in figure 6.7. Here also we have considered subjects' results with validation scores of 10 or more. The preferences for no smoothing and linear smoothing are higher compared to Kalman smoothing. Overall, linear smoothing is preferred more times. The preferences for smoothing methods after removing invalid blocks are presented in figure 6.8. In this case also we have observed no smoothing and linear smoothing is preferred more times than Kalman smoothing; overall linear smoothing has more preferences.

Validation	*No smoothing vs Linear*		*Linear vs Kalman*		*Kalman vs No smoothing*	
cutoff	t	α	t	α	t	α
1	-0.8623	> 0.2	4.1514	**0.01**	-8.4652	**0.01**
2	-0.8623	> 0.2	4.1514	**0.01**	-8.4652	**0.01**
3	-0.8623	> 0.2	4.1514	**0.01**	-8.4652	**0.01**
4	-0.8930	> 0.2	4.2654	**0.01**	-8.4936	**0.01**
5	-0.8337	> 0.2	4.0598	**0.01**	-7.7575	**0.01**
6	-0.8337	> 0.2	4.0598	**0.01**	-7.7575	**0.01**
7	-0.6081	> 0.2	3.9109	**0.01**	-7.0384	**0.01**
8	-1.2524	> 0.2	4.3299	**0.01**	-5.9982	**0.01**
9	-0.5649	> 0.2	4.7932	**0.01**	-6.4502	**0.01**
10	-0.4065	> 0.2	6.0472	**0.01**	-6.8313	**0.01**
11	-0.1584	> 0.2	5.1335	**0.01**	-4.6508	**0.01**
12	1.3416	> 0.2	2.6405	0.05	-3.2160	0.02
13	0.5000	> 0.2	1.7300	0.20	-2.5152	0.05
14	0.2047	> 0.2	1.1062	> 0.2	-1.5901	0.20
15	-0.6070	> 0.2	0.1876	> 0.2	0.3568	> 0.2

Table 6.5: Paired t-test statistics for the smoothing methods

We present paired t-test statistics for three smoothing comparisons in table 6.5 for different validation cutoffs (after removing invalid blocks). The preference for no

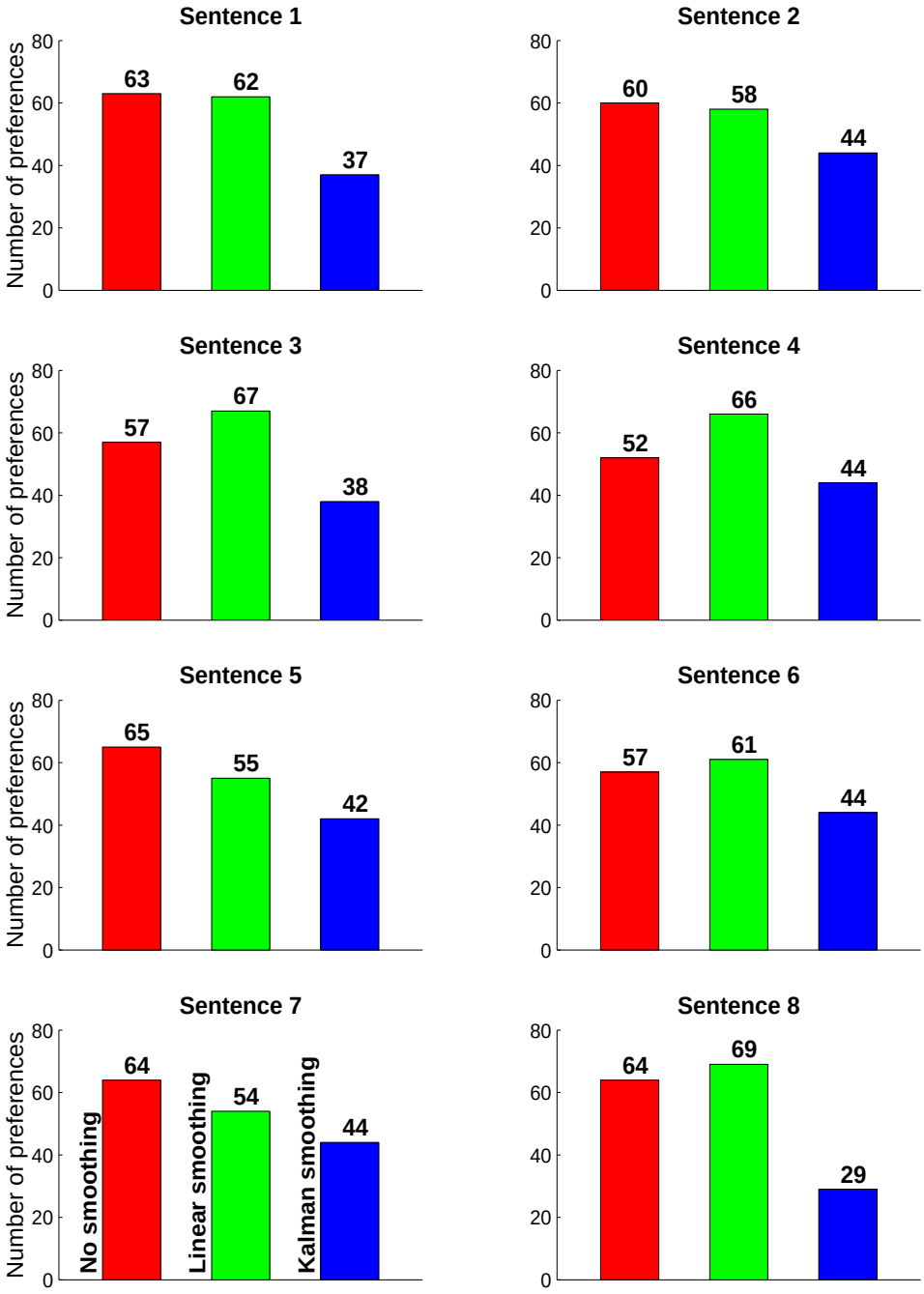

Figure 6.7: Smoothing evaluation, validation cutoff 10

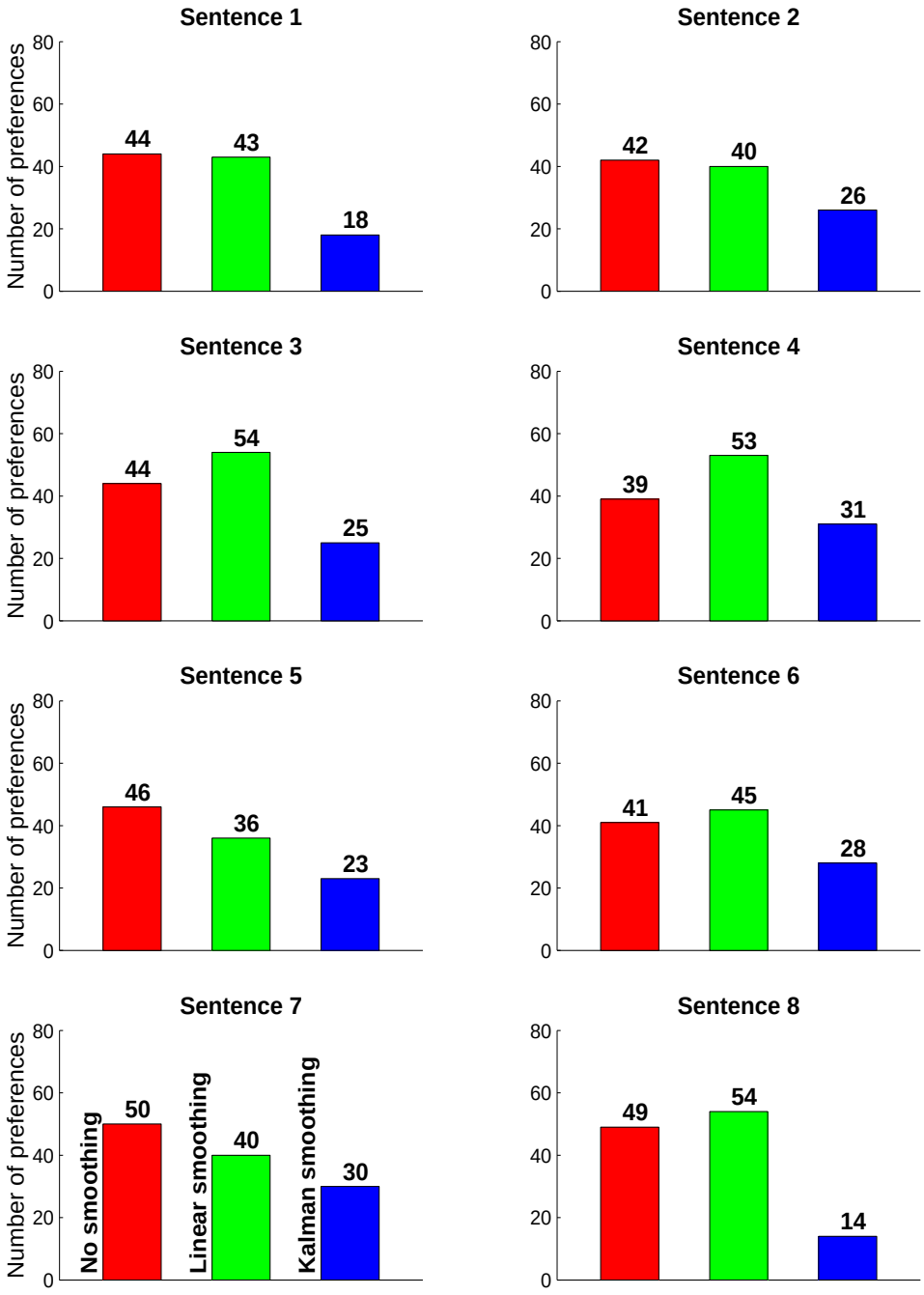

Figure 6.8: Smoothing evaluation, validation cutoff 10 plus block validation check (after removing invalid blocks)

smoothing over linear smoothing is not statistically significant. However there is a significant preference towards linear smoothing over Kalman smoothing except for high validation cutoffs, where it is not significant. Similarly, the preference for no smoothing over Kalman smoothing is significant, but for high validation cutoffs it is less significant.

6.4.3 Kalman-Kalman vs LSF-linear

The preferences for Kalman join cost with Kalman smoothing compared to LSF join cost with linear smoothing are shown in figure 6.9. LSF-linear is preferred more times than Kalman-Kalman in all sentences. The statistical results in table 6.6 also conclude that the preference towards LSF-linear is significant.

Validation score	*Kalman-Kalman vs LSF-linear*	
	t	α
1	-7.5245	**0.01**
2	-7.5245	**0.01**
3	-7.5245	**0.01**
4	-7.6502	**0.01**
5	-7.5348	**0.01**
6	-7.5348	**0.01**
7	-7.7497	**0.01**
8	-8.0958	**0.01**
9	-8.7794	**0.01**
10	-9.6776	**0.01**
11	-8.7767	**0.01**
12	-5.9161	**0.01**
13	-7.2022	**0.01**
14	-3.9886	**0.01**
15	N/A	N/A

Table 6.6: Paired t-test statistics for the Kalman-Kalman and LSF-linear comparison

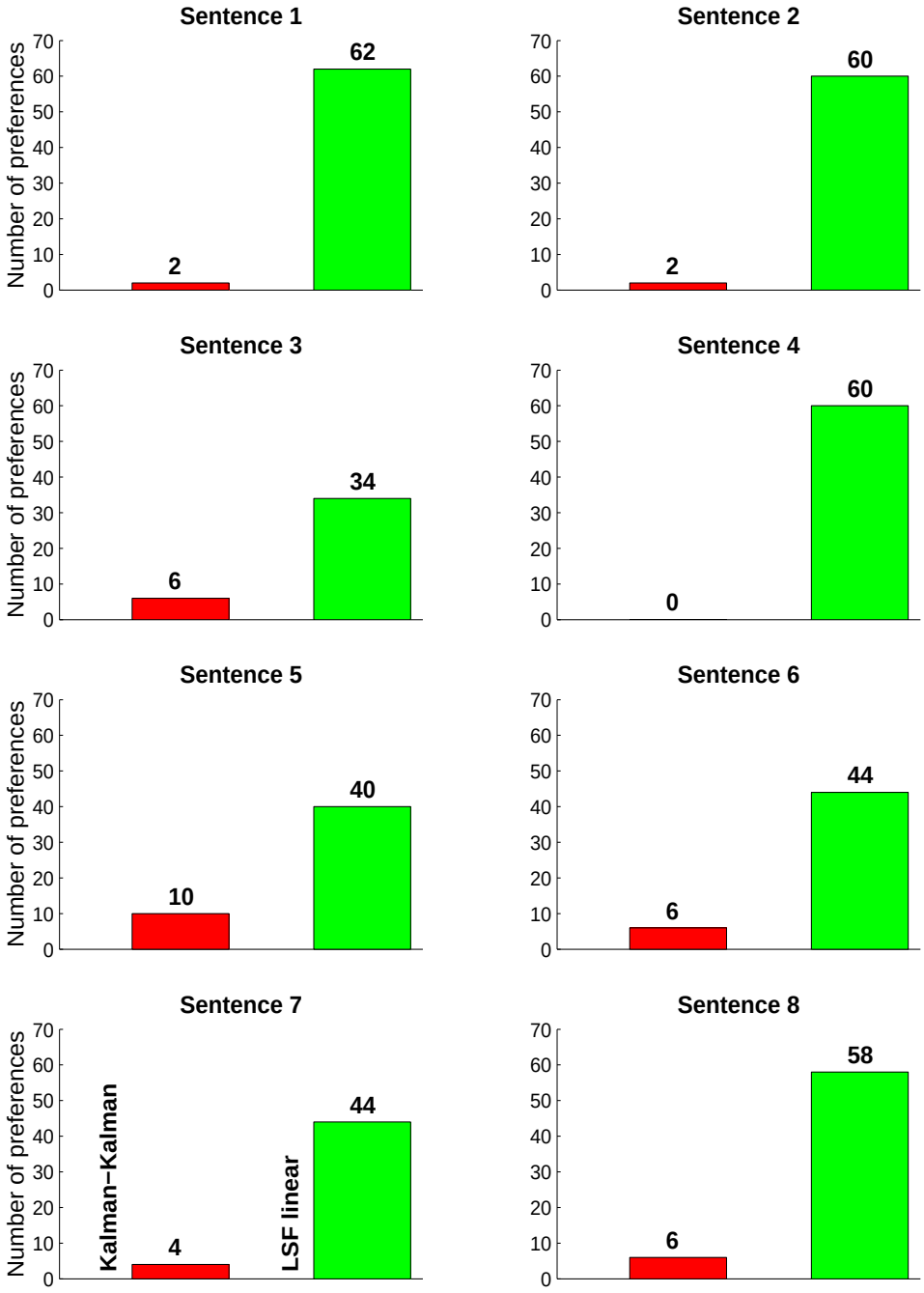

Figure 6.9: Kalman-Kalman and LSF linear comparison, validation cutoff 10

6.5 Conclusions

In this chapter, the three best spectral distances from previous chapters and different smoothing methods were evaluated by conducting a listening test. The join cost functions were:

- *Mahalanobis distance on LSFs with their deltas of the frames at the join,*
- *Mahalanobis distance computed using MCA coefficients of seven frames*
- *Kalman-filter based join cost*

The linear smoothing technique, smoothed parameters (LSFs) obtained from Kalman filter and no smoothing were compared in the listening test. In addition to these, both join cost and smoothing using Kalman filter was compared with LSF join cost and linear smoothing.

The results from the listening test indicated that LSF join cost has more preferences than MCA join cost and Kalman join cost. These results reconfirmed our previous perceptual test results (refer table 6.1). Though the LSF join cost has more preferences, the preference for it over MCA join cost is not statistically significant. The preference towards MCA join cost over Kalman join cost is also not statistically significant. For low validation cutoffs, LSF join cost preference over Kalman join cost is statistically significant. But, for high validation cutoffs (more consistent subjective results) it is less significant.

Linear smoothing was preferred more times than no smoothing and Kalman smoothing. There is no significant preference between no smoothing and linear smoothing. However, the preference for both of them over Kalman smoothing is significant except for high validation cutoffs, where the significance is lower. The preference for LSF join cost and linear smoothing over Kalman join cost and Kalman smoothing is statistically significant.

Unfortunately, Kalman filtering did not perform well in both join cost and smoothing. This may be due to inadequate learning of the hidden state space by the model. Another study (Adell 2003) observed a similar effect, i.e. the state parameters are ineffective and the process is dominated by the observation noise. One way of alleviating this problem

might be by training the models using articulatory data so that they learn the hidden state correctly. Another way might be controlling the observation noise.

Chapter 7

Conclusion

The main objective of this thesis was to formulate a standard method to design and evaluate join cost functions. A join cost is computed using a *join cost function*, which uses a *distance measure* on some *parameterisation* of the speech signal. The ideal join cost is one that correlates highly with human listeners' perception of discontinuity at unit concatenation points (joins). The derivation and evaluation of a join cost function which meets this requirement is a very difficult task.

In this thesis, we investigated various join cost functions based on spectral distance measures and evaluated them using a perceptual listening test. This test used synthetic examples (using a male north-American voice) of **whole (continuous) sentences with polysyllabic words** generated using an experimental version of the state-of-the-art unit selection synthesis system, *rVoice*. The spectral representations used were: Mel-frequency cepstral coefficients (MFCC), line spectral frequencies (LSF), Multiple centroid analysis coefficients (formant-like parameters) and LPC spectra.

Several distance measures, which can be used in join cost functions, were compared based on their correlations with mean listener scores. We proposed join cost functions based on a learned underlying representation of speech signal. The linear dynamical model (sometimes known as **Kalman filter**) was used to learn this representation. This model also able to smooth the observations. We also investigated using linear dynamical models for join cost computation and smoothing in a single operation.

Finally, we evaluated the best three join cost functions subjectively by conducting

another listening test, where stimuli were synthesised using a female north-American voice. In this test, we also compared different smoothing methods.

We start this chapter with an analysis of the perceptual listening test and our evaluation tool (correlation). Then we summarise our correlation results for various join cost functions in section 7.2. In section 7.3, we summarise our subjective evaluation. Finally we conclude this chapter with few pointers to future work.

7.1 Perceptual test paradigm

Since the design and evaluation of a join cost function is an iterative process, it is not usually practical to conduct perceptual listening tests for each and every proposed function. A more efficient method is to conduct a single perceptual listening test covering wide range of synthetic stimuli. From this, we obtain listeners' ratings and design the join cost function to maximise correlations with those ratings.

Many previous studies (table 3.1 in page 46) used a similar methodology but used mono-syllabic isolated words for constructing synthetic stimuli. In this thesis, we proposed a variation on the usual perceptual test paradigm by deliberately including a **wide range of qualities** of join in **polysyllabic words (continuous speech)**. Donovan (2001) also allowed the occurrence of spectral discontinuities in the stimuli, but his stimuli is based on CV pairs, If we use natural sentences as our stimuli, the effects of co-articulation and wider contexts can be taken into account and better objective measures can be formulated.

We designed a perceptual listening test to measure the degree of perceived concatenation discontinuity in whole (continuous speech) sentences generated by the experimental version of a state-of-the-art speech synthesis system, *rVoice*. We chose five American diphthongs as these have more prominent spectral discontinuities in the middle due to moving formant values.

Since the stimuli were created by a typical unit selection synthesiser, we believe that the results we obtain on such stimuli apply to:

- Any speech produced by the same synthesiser, not just American English diphthongs

- Any other unit selection synthesiser.

7.2 Summary of join cost results

Usually the join cost function is derived from a few sub-costs defined on spectral and prosodic properties of the speech frames on either side of a join between two units. These sub costs are, for example: *difference in energy*, *difference in* $F0$ and *mismatch in spectral parameters*.

The energy and $F0$ differences are rather easy to compute since these parameters are scalars. We can represent them in some perceptual scale (logarithmic scale for energy and Bark or Mel scale for $F0$) to get better costs. However, to achieve a good spectral cost (mismatch in spectral parameters) is a hard problem due to many possible speech parameterisations and distance metrics. Hence, we focused only on the spectral component of the join cost function and omitted energy and $F0$ components. Since we used only the spectral component we termed the join costs obtained with our join cost functions "spectral distances".

7.2.1 Correlation as an evaluation tool

Since we have a 1-5 scale for rating the discontinuities, correlation is used as an evaluation tool. First, we computed correlations of individual perceptual data (i.e. each of our ten cases) with various distance measures. We assumed that these ten cases were independent and calculated the 1% statistical significance ($\alpha = 0.01$) of correlations for each case. Then various join cost functions compared based on the number of 1% significant correlations achieved out of ten cases. We also used global correlation coefficient for comparison purposes, which is computed by pooling all the perceptual data.

7.2.2 Single frame spectral distances

The single frame distances were computed using parameterisations of frames on either side of the join. We experimented with different parameterisations – Mel frequency cepstral coefficients (MFCCs), line spectral frequencies (LSFs), multiple centroid analysis

(MCA) coefficients and LPC spectra – and distance metrics: Euclidean, absolute, Mahalanobis and symmetric Kullback-Leibler (K-L) (also Itakura-Saito and COSH distances for LPC spectra).

In table 3.8 we presented the number of 1% significant correlations and global correlation for various distance metrics computed using various speech parameterisations (MFCC, LSF and MCA coefficients). It is observed that Mahalanobis distance using LSFs plus delta coefficients has the most (**six**) 1% significant correlations. The other good predictors are: Mahalanobis distance using MFCCs and Mahalanobis distance on MCA parameters with their deltas. They both got **five** 1% significant correlations. The global correlation values are high for K-L and Mahalanobis distances computed using LSFs.

The distances computed using LPC spectra yielded **three** significant correlations in contradiction to past findings (Klabbers & Veldhuis 1998, Founda et al. 2001), where they found K-L distance on LPC spectra has high prediction rates. This may be due to different testing paradigms: those studies were based on diphone synthesis (not unit selection) of languages other than English and they used single words (syllables).

If we consider the storage requirements of the inventory, i.e. space to store all these parameters for join frames, then Mahalanobis distance using MCA plus deltas is a reasonable choice, since their size (including deltas) is 12 compared to 26 for MFCCs and 24 for LSFs.

7.2.3 ANOVA on single-frame distances

We have also performed several ANOVA tests on single-frame spectral distances and data obtained from perceptual listening experiments. The main observations from these tests are as follows:

- The four distance measures are significantly different from one another for a given spectral feature
- The effect of distance measure on the spectral distance is much stronger that that due to spectral feature

- The spectral distance relate to sentences or diphthongs, spectral features and distance measures
- The spectral distance do not relate to whether it is the first or second sentence in the given diphthong
- The listeners' scores differ significantly from one sentence or diphthong to another and also one listener to another, but the overall listeners' F value is lower than the spectral features F value and the distance measures F value
- The test stimuli are significantly different for given diphthong

7.2.4 Multi-frame spectral distances

We also computed spectral distances using multiple frames of the two segments on either side of the join. This is to include the effect of neighbouring frames of join frames on discontinuity. We derived different distances – three-frame, five-frame, seven-frame and nine-frame – based on number of frames used, for example, a three frame distance is obtained using middle frame at the join and one frame on either side of the join. To calculate a distance, we average distances of all these frames, hence we also call them "average distances".

It is evident from our study that single frame distances result in better correlations than those of multi-frame distances in most cases; also, the storage requirements for single frame distances are low. Only distances computed using MCAs have a greater number of 1% significant correlations: we achieved **seven** 1% significant correlations for Mahalanobis distance using MCAs for seven and nine frame distance (tables 3.17 and 3.21). We present the number of 1% significant correlations and global correlation values for distances computed using seven frames in table 3.18. The global correlations are high in the case of Mahalanobis and K-L distances.

But to use the above distances as join cost functions, one has to pay a price in terms of substantial disk usage, i.e., a seven fold increase in storage requirements compared to single frame distances, as we need to store fourteen frames for each segment to compute spectral distance measures.

7.2.5 Weighted spectral distances

In order to achieve a higher number of 1% significant correlations in our ten sentences (as listed in table 3.2 in page 48), we constructed new spectral distance measures as weighted sums of individual distances. The weights were computed using the standard least squares method by fitting a linear model with and without intercept to perceptual data. We carried out both closed set and open set experiments for computing weights. In open set experiments, training and test sets were chosen using a cross-validation procedure. We also used triangular weighting for multi-frame distances, i.e. providing higher weights to frames at the join and lower weights to frames away from the join.

7.2.5.1 Weighted combinations of single-frame distances

A weighted sum of distances obtained from MFCC, LSF and MCA coefficients, produced marginal improvements in correlations for some cases. We achieved **eight** (our highest number) 1% significant correlations for distances computed by weighting individual MFCCs and their deltas (table 4.9) and also in the case of LSFs plus delta coefficients (table 4.11). Unfortunately, these results may not generalise, since weights for the above cases were computed in a closed set experiment. In open set experiments fewer 1% significant correlations were observed, the highest being **five** for the LSFs with their deltas in *open set(5+5)* (table 4.11).

7.2.5.2 Weighted multi-frame distances

In multi-frame distances, the triangular weights gave better results compared to their least squares counterpart. But there was not much improvement compared to average multi-frame distances.

Overall, the weighting did not result in much improvement, and of course one has to pay a price for much higher storage requirements since multiple parameterisations of the speech signal must be stored.

7.2.6 Kalman filter-based join cost

To include co-articulation effects to some extent, we trained linear dynamical models (LDM) for each phone to model line spectral frequencies (LSF) obtained from natural speech.This model uses an underlaying hidden state space in which it makes smooth and continuous trajectories. The Expectation-Maximisation (EM) algorithm used in training is sensitive to initial values. So we experimented with three initialisation schemes: *AR(1): first order auto-regressive process*, *Factor analysis* and *Empirical values*.

The objective join cost function was defined based on the log likelihood of the observations given the model. Based on the shape of these likelihood plots, three different analytical measures were derived. Then, correlation values were computed between these measures and mean listener scores. The main observations from these results are:

- Initialising the model with *empirical values* resulted more 1% significant correlations
- The first measure is best among our three analytical measures, which uses a simple average of the log likelihood over 5 frames centered on the join (figure 5.6 in page 116)
- There is no improvement observed using models trained on Bark-scaled LSFs; their performance is poor compared to models trained on original LSFs.

7.3 Summary of Subjective Evaluation

Three of the best spectral distances found in our study, along with various smoothing methods were evaluated in a listening test.

7.3.1 Join cost functions

The join cost functions evaluated using the listening test were: *Mahalanobis distance on LSFs, with their deltas, of the frames at the join*, *Mahalanobis distance on MCA coefficients of seven frames* and *Kalman filter-based join cost*.

The results from the listening test (number of preferences) and paired t-test statistics (significance level) are shown in table 7.1. These results can be summarised as follows:

Validation cutoff	No. of preferences			Significance level (α)		
	LSF	MCA	Kalman	LSF-MCA	MCA-Kalman	LSF-Kalman
10	494	340	198	0.20	> 0.20	**0.01**
11	273	195	108	0.20	0.20	0.02
12	124	83	27	0.10	0.10	0.02
13	46	38	6	> 0.20	0.10	0.05
14	28	24	5	> 0.20	0.10	0.05
15	16	12	5	> 0.20	> 0.20	0.10

Table 7.1: Join cost evaluation for different validation cutoffs after removing invalid blocks (for all validation cutoffs refer to table 6.4)

- LSF join cost has more preferences over MCA join cost and over Kalman join cost
- For low validation cutoffs, LSF join cost preference over Kalman join cost is statistically significant

7.3.2 Smoothing methods

No smoothing, linear smoothing method and smoothed parameters (LSFs) obtained from Kalman filter were compared in the listening test. In table 7.2, we present the number of preferences and significance levels of three smoothing methods for different validation cutoffs.

- Linear smoothing and no smoothing were preferred over Kalman smoothing.
- There is no significant preference between no smoothing and linear smoothing.
- The preference for both of them over Kalman smoothing is significant

In addition to this, both join cost and smoothing using Kalman filter was compared to LSF join cost and linear smoothing. The preference for LSF join cost and linear smoothing over Kalman join cost and smoothing is large and statistically significant.

Validation cutoff	No. of preferences			Significance level (α)		
	No	Linear	Kalman	No-linear	Linear-Kalman	No-Kalman
10	355	365	195	> 0.20	**0.01**	**0.01**
11	198	201	105	> 0.20	**0.01**	**0.01**
12	95	77	38	> 0.20	0.05	0.02
13	36	32	10	> 0.20	0.20	0.05
14	21	20	10	> 0.20	> 0.20	0.20
15	6	8	7	> 0.20	> 0.20	> 0.20

Table 7.2: Smoothing evaluation for different validation cutoffs after removing invalid blocks (for all validation cutoffs refer to table 6.5)

7.4 Contributions of this thesis

This thesis provides a method for designing and evaluation of join cost functions in unit selection using a single perceptual test paradigm. In this test we deliberately included a **wide range of qualities** of join in natural (continuous) speech. A similar perceptual test with some modifications could also be used to design and evaluate target cost functions, where the test stimuli may consist of different contexts of the target units.

We investigated several spectral distance-based join cost functions from our subjective evaluations. This thesis also provides more insight into the spectral component of the join cost function, mainly how various speech parameterisations perform in detecting concatenation discontinuities and their performance in single frame and multi-frame distances. In addition, we investigated different ways of combining various speech parameterisations and different distance metrics.

In this thesis, we observed that Kalman filter-based join cost did not work, at least in our implementation. The simple[1] spectral distance measure, **Mahalanobis distance using LSFs plus deltas** was found to be the best predictor overall.

To test the performance of Kalman filter smoothing we compared this method with linear smoothing and no smoothing. We observed **linear smoothing** performs better and Kalman filter smoothing has worst performance. These results indicated that linear

[1] Occupies less disk space as well as lower computational cost

smoothing can improve the quality of synthetic speech but one should be very careful to employ any kind of smoothing as it sometimes produce artifacts and thus deteriorates the quality.

7.5 Future work

The most obvious extension to our work would be to make the Kalman filter work properly for join cost and smoothing. From our experience, we feel our models did not learn the hidden state space properly and thus the hidden state does not, in fact, model underlying articulator movements in speech production. Other studies related to tying parameters for LDMs (Adell 2003) also observed that the observation noise is doing all the work and the state is not doing much. Hence it may be worthwhile to train the models using articulatory data instead of spectral parameters to learn the hidden state correctly.

We chose diphthongs for our first perceptual test, since these are very difficult to join. Our hypothesis is that if our join cost function works for diphthongs then it will performs well in other phones too. In our study, we observed that the performance of our best join cost functions was good in most of our ten cases, though it varies among these cases. It would be worthwhile to test these join cost functions in other phone categories. In addition, for some phones (e.g. fricatives), the join cost function may not be important because of their "noisy" nature, so the joins would be masked and we do not hear them. Hence, it would be useful to investigate phone-specific (at least phone-class specific) join cost functions.

As we mentioned before, our join cost functions contain the spectral component only. This is because our goal was to analyse the performance of various spectral distance measures for use in the join cost function. However, we strongly believe that adding good $F0$ and energy costs to our best join cost function will increases its performance; this was beyond the scope of this thesis.

Appendix A

rVoice Text-to-Speech System

rVoice is a general purpose text-to-speech system that converts any type of text into spoken language. It produces speech that is sometimes indistinguishable from human voice in terms of naturalness, intelligibility and accuracy and is believed to be of very high quality TTS system[1]. rVoice uses latest *unit selection* concatenative technology for synthesised speech. Currently, rVoice offers a number of different voices in English (UK, US, Australian), German, Greek and American Spanish.

A.1 Voices used in this research

We were provided with two English voices by Rhetorical systems Ltd. for this research; both are north-American voices. The first one is an adult male voice (US_M009) and the other one is a female voice (US_F006). Interactive synthesis examples using the male voice are available on Rhetorical website (*http://www.rhetorical.com/cgi-bin/demo.cgi*)

A.2 Advantages of using rVoice

- Offers high quality both in terms of naturalness and intelligibility
- Availability of good voices
- Flexibility in plugging various join cost functions

[1] From rVoice spec-sheet at Rhetorical website, http://www.rhetoricalsystems.com/pdf/specsheet42.pdf

A.3 Limitations in using rVoice

- Limited access to voices; for male voice, we do not have access to wave files. Hence, we were not able to use pitch-synchronous frames for our distance measures.
- Limited access to source code, only to join cost functions. So we had to re-implement waveform generation.

Despite the above limitations, we preferred *rVoice* instead of Festival because of its advantages. Though we do not have access to male voice wave files, it may even be preferable to use a different voice for our second listening test. Fortunately, Rhetorical Systems Ltd. provided us a female voice with wave files. This allowed us to use different smoothing techniques by implementing the residual excited linear prediction (RELP) synthesis method for waveform generation.

Appendix B

Weighted Multi-frame Distances

B.1 Weighted three frame distances

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	0.31	0.34	0.28	0.33	0.22	0.34
	0.69	**0.67**	**0.71**	**0.70**	**0.76**	**0.63**
ow	0.21	0.20	0.22	0.20	0.22	0.21
	0.55	**0.53**	**0.54**	**0.53**	**0.55**	**0.53**
ay	0.31	0.27	0.27	0.18	0.39	0.22
	0.72	**0.76**	**0.68**	**0.76**	**0.65**	**0.76**
aw	0.29	0.32	0.27	0.33	0.18	0.34
	0.81	**0.81**	**0.82**	**0.82**	**0.79**	**0.80**
oy	0.03	0.06	0.03	0.08	0.20	0.28
	0.17	0.19	0.14	0.17	0.21	0.25

Table B.1: Correlation between perceptual scores and weighted distance measures computed using **three** frames of **MFCC**s in **open set (9+1)** experiment

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	0.25	0.25	0.22	0.26	0.20	0.32
	0.70	**0.65**	**0.73**	**0.67**	**0.76**	**0.54**
ow	0.16	0.15	0.23	0.16	0.22	0.21
	0.52	**0.45**	**0.56**	**0.48**	**0.53**	0.39
ay	0.22	0.16	0.16	0.03	0.40	0.16
	0.71	**0.75**	**0.62**	**0.72**	**0.63**	**0.75**
aw	0.27	0.32	0.26	0.32	0.19	0.36
	0.82	**0.81**	**0.78**	**0.81**	**0.78**	**0.78**
oy	0.03	0.06	-0.00	0.03	0.20	0.27
	0.17	0.20	0.14	0.18	0.19	0.25

Table B.2: Correlation between perceptual scores and weighted distance measures computed using **three** frames of **MFCC**s in **open set (9+1)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	0.17	0.19	0.14	0.18	0.21	0.26
	0.32	0.35	0.30	0.36	0.23	0.36
ow	**0.67**	**0.65**	**0.69**	**0.68**	**0.75**	**0.61**
	0.19	0.19	0.20	0.19	0.22	0.21
ay	**0.55**	**0.53**	**0.54**	**0.53**	**0.55**	**0.50**
	0.31	0.27	0.26	0.18	0.40	0.22
aw	**0.72**	**0.76**	**0.68**	**0.76**	**0.66**	**0.76**
	0.29	0.32	0.27	0.33	0.18	0.34
oy	**0.81**	**0.81**	**0.82**	**0.82**	**0.79**	**0.80**
	0.02	0.06	0.04	0.07	0.20	0.28

Table B.3: Correlation between perceptual scores and weighted distance measures computed using **three** frames of **MFCC**s in **open set (8+2)** experiment

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	0.17	0.20	0.14	0.20	0.19	0.28
	0.28	0.31	0.27	0.34	0.22	0.37
ow	**0.60**	**0.55**	**0.68**	**0.58**	**0.72**	**0.48**
	0.15	0.14	0.20	0.16	0.22	0.21
ay	**0.51**	0.41	**0.56**	**0.44**	**0.53**	0.33
	0.21	0.14	0.15	0.02	0.38	0.13
aw	**0.72**	**0.75**	**0.62**	**0.72**	**0.65**	**0.73**
	0.28	0.32	0.26	0.31	0.20	0.35
oy	**0.81**	**0.81**	**0.78**	**0.80**	**0.79**	**0.78**
	0.04	0.06	-0.00	0.01	0.22	0.26

Table B.4: Correlation between perceptual scores and weighted distance measures computed using **three** frames of **MFCC**s in **open set (8+2)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	0.30	0.31	0.26	0.31	0.20	0.32
	0.18	0.17	0.21	0.18	0.20	0.17
ow	0.26	0.21	0.19	0.09	0.32	0.16
	0.27	0.31	0.23	0.31	0.14	0.36
ay	0.02	0.05	-0.00	0.04	0.15	0.24
	0.66	**0.63**	**0.69**	**0.63**	**0.71**	**0.57**
aw	**0.52**	**0.52**	**0.51**	**0.52**	**0.55**	**0.56**
	0.66	**0.69**	**0.66**	**0.74**	**0.59**	**0.67**
oy	**0.76**	**0.77**	**0.73**	**0.77**	**0.75**	**0.81**
	0.16	0.17	0.13	0.16	0.19	0.23

Table B.5: Correlation between perceptual scores and weighted distance measures computed using **three** frames of **MFCC**s in **open set (5+5)** experiment

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	0.24	0.22	0.18	0.19	0.16	0.23
	0.12	0.10	0.17	0.11	0.18	0.10
ow	0.15	0.09	0.06	-0.06	0.27	0.07
	0.23	0.29	0.18	0.28	0.10	0.37
ay	0.01	0.05	-0.05	-0.01	0.11	0.18
	0.59	**0.61**	**0.54**	**0.56**	**0.54**	0.38
aw	**0.53**	**0.52**	**0.52**	**0.49**	**0.53**	**0.47**
	0.58	**0.66**	**0.56**	**0.72**	0.45	**0.59**
oy	**0.72**	**0.76**	**0.62**	**0.77**	**0.68**	**0.78**
	0.14	0.17	0.10	0.13	0.14	0.19

Table B.6: Correlation between perceptual scores and weighted distance measures computed using **three** frames of **MFCC**s in **open set (5+5)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	0.32	0.35	0.29	0.35	0.22	0.36
	0.70	**0.68**	**0.72**	**0.71**	**0.76**	**0.63**
ow	0.22	0.21	0.23	0.20	0.23	0.20
	0.54	**0.53**	**0.53**	**0.52**	**0.54**	**0.52**
ay	0.32	0.29	0.28	0.21	0.40	0.25
	0.72	**0.76**	**0.69**	**0.76**	**0.66**	**0.76**
aw	0.29	0.32	0.27	0.33	0.20	0.34
	0.81	**0.81**	**0.82**	**0.82**	**0.79**	**0.80**
oy	0.03	0.06	0.04	0.08	0.20	0.28
	0.19	0.22	0.17	0.21	0.23	0.28

Table B.7: Correlation between perceptual scores and weighted distance measures computed using **three** frames of **MFCC**s in **closed set** experiment

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	0.28	0.30	0.24	0.32	0.22	0.37
	0.72	**0.68**	**0.73**	**0.72**	**0.77**	**0.58**
ow	0.21	0.18	0.28	0.18	0.25	0.20
	0.56	**0.51**	**0.57**	**0.50**	**0.56**	**0.45**
ay	0.25	0.20	0.20	0.09	0.41	0.21
	0.72	**0.76**	**0.64**	**0.73**	**0.64**	**0.76**
aw	0.28	0.31	0.27	0.32	0.21	0.36
	0.82	**0.81**	**0.83**	**0.83**	**0.79**	**0.78**
oy	0.03	0.06	0.00	0.04	0.21	0.29
	0.20	0.25	0.16	0.26	0.21	0.30

Table B.8: Correlation between perceptual scores and weighted distance measures computed using **three** frames of **MFCC**s in **closed set** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.05	0.15	0.11	0.27	0.30	0.15	0.39
	0.74	**0.72**	**0.76**	**0.73**	**0.71**	**0.72**	**0.62**
ow	0.32	0.29	0.31	0.23	0.26	0.40	0.22
	0.51	**0.46**	0.44	0.39	**0.46**	0.41	0.19
ay	0.12	0.08	0.07	0.01	0.37	0.13	-0.00
	0.52	**0.67**	**0.52**	**0.70**	**0.67**	**0.48**	**0.51**
aw	0.29	0.38	0.22	0.37	0.22	0.34	**0.53**
	0.79	**0.79**	**0.79**	**0.77**	**0.80**	**0.78**	**0.69**
oy	0.26	0.25	0.21	0.21	0.18	0.31	0.44
	0.19	0.27	0.21	0.28	0.37	0.26	0.39

Table B.9: Correlation between perceptual scores and weighted distance measures computed using **three** frames of **LSF**s in **open set (9+1)** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.04	0.11	0.12	0.24	0.31	0.14	0.32
	0.71	**0.65**	**0.72**	**0.68**	**0.68**	**0.70**	**0.58**
ow	0.32	0.22	0.31	0.14	0.28	0.39	0.16
	0.50	0.43	0.40	0.31	0.41	0.40	0.19
ay	0.17	0.15	0.09	0.07	0.39	0.15	0.03
	0.48	**0.51**	**0.52**	**0.58**	**0.65**	**0.49**	0.44
aw	0.19	0.32	0.10	0.39	0.24	0.26	**0.51**
	0.79	**0.78**	**0.79**	**0.77**	**0.79**	**0.79**	**0.67**
oy	0.14	0.20	0.10	0.12	0.16	0.25	0.35
	0.15	0.29	0.19	0.29	0.35	0.25	0.41

Table B.10: Correlation between perceptual scores and weighted distance measures computed using **three** frames of **LSF**s in **open set (9+1)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.18	0.27	0.20	0.28	0.37	0.26	0.39
	0.05	0.17	0.12	0.30	0.30	0.15	0.42
ow	**0.74**	**0.72**	**0.75**	**0.73**	**0.71**	**0.71**	**0.62**
	0.33	0.28	0.31	0.22	0.27	0.40	0.21
ay	**0.52**	**0.48**	**0.45**	**0.42**	**0.47**	0.41	0.27
	0.13	0.10	0.08	0.03	0.38	0.14	0.00
aw	**0.52**	**0.65**	**0.52**	**0.69**	**0.67**	**0.49**	**0.50**
	0.29	0.38	0.22	0.37	0.23	0.36	**0.55**
oy	**0.79**	**0.79**	**0.79**	**0.77**	**0.80**	**0.78**	**0.69**
	0.27	0.25	0.21	0.22	0.18	0.30	0.44

Table B.11: Correlation between perceptual scores and weighted distance measures computed using **three** frames of **LSF**s in **open set (8+2)** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.15	0.29	0.18	0.29	0.35	0.23	0.42
	0.05	0.15	0.11	0.29	0.31	0.14	0.39
ow	**0.74**	**0.67**	**0.72**	**0.68**	**0.69**	**0.71**	**0.59**
	0.35	0.23	0.32	0.14	0.30	0.40	0.16
ay	**0.49**	0.37	0.40	0.29	0.40	0.40	0.26
	0.19	0.23	0.10	0.18	0.40	0.16	0.08
aw	0.44	**0.47**	**0.50**	**0.60**	**0.64**	**0.48**	0.44
	0.15	0.27	0.07	0.39	0.23	0.24	**0.49**
oy	**0.77**	**0.77**	**0.77**	**0.75**	**0.79**	**0.78**	**0.66**
	0.08	0.16	0.04	0.08	0.15	0.22	0.30

Table B.12: Correlation between perceptual scores and weighted distance measures computed using **three** frames of **LSF**s in **open set (8+2)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.04	0.12	0.10	0.23	0.28	0.15	0.33
	0.27	0.18	0.29	0.13	0.19	0.39	0.06
ow	0.08	0.02	0.04	-0.04	0.34	0.13	-0.05
	0.24	0.36	0.19	0.34	0.14	0.31	**0.56**
ay	0.25	0.23	0.23	0.17	0.18	0.31	0.29
	0.71	**0.69**	**0.73**	**0.70**	**0.69**	**0.71**	**0.58**
aw	**0.47**	0.32	0.40	0.25	0.43	0.38	0.14
	0.52	**0.66**	**0.52**	**0.69**	**0.66**	**0.48**	**0.52**
oy	**0.79**	**0.79**	**0.79**	**0.76**	**0.80**	**0.77**	**0.70**
	0.19	0.28	0.22	0.29	0.38	0.28	0.41

Table B.13: Correlation between perceptual scores and weighted distance measures computed using **three** frames of **LSF**s in **open set (5+5)** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.02	0.05	0.09	0.13	0.27	0.14	0.22
	0.23	0.05	0.30	-0.05	0.18	0.39	-0.03
ow	0.06	-0.03	0.04	-0.10	0.33	0.16	-0.07
	0.16	0.26	0.08	0.29	0.13	0.26	**0.53**
ay	0.20	0.17	0.20	0.08	0.18	0.29	0.20
	0.65	**0.65**	**0.68**	**0.65**	**0.65**	**0.70**	**0.57**
aw	**0.46**	0.30	0.35	0.17	0.35	0.39	0.12
	0.49	**0.62**	**0.51**	**0.65**	**0.62**	**0.49**	**0.50**
oy	**0.78**	**0.78**	**0.78**	**0.75**	**0.78**	**0.78**	**0.70**
	0.15	0.28	0.24	0.29	0.35	0.28	0.41

Table B.14: Correlation between perceptual scores and weighted distance measures computed using **three** frames of **LSF**s in **open set (5+5)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.05	0.16	0.12	0.29	0.30	0.15	0.39
	0.74	**0.72**	**0.76**	**0.73**	**0.72**	**0.72**	**0.62**
ow	0.33	0.30	0.31	0.24	0.29	0.39	0.23
	0.50	0.44	0.43	0.39	**0.47**	0.41	0.21
ay	0.13	0.11	0.07	0.05	0.38	0.12	0.06
	0.52	**0.67**	**0.52**	**0.71**	**0.67**	**0.49**	**0.56**
aw	0.29	0.38	0.22	0.36	0.23	0.35	**0.55**
	0.79	**0.79**	**0.79**	**0.77**	**0.80**	**0.78**	**0.68**
oy	0.26	0.25	0.20	0.21	0.18	0.30	0.44
	0.18	0.28	0.20	0.29	0.37	0.27	0.41

Table B.15: Correlation between perceptual scores and weighted distance measures computed using **three** frames of **LSF**s in **closed set** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.05	0.13	0.14	0.27	0.32	0.15	0.34
	0.72	**0.66**	**0.74**	**0.69**	**0.70**	**0.72**	**0.59**
ow	0.41	0.29	0.33	0.20	0.35	0.40	0.23
	0.52	**0.56**	0.44	0.43	0.43	0.41	0.33
ay	0.18	0.14	0.11	0.09	0.39	0.16	0.08
	0.50	**0.57**	**0.52**	**0.64**	**0.66**	**0.49**	**0.49**
aw	0.22	0.34	0.18	0.38	0.26	0.30	**0.57**
	0.79	**0.79**	**0.79**	**0.77**	**0.80**	**0.79**	**0.68**
oy	0.18	0.20	0.15	0.14	0.17	0.29	0.36
	0.13	0.30	0.20	0.31	0.36	0.26	**0.46**

Table B.16: Correlation between perceptual scores and weighted distance measures computed using **three** frames of **LSF**s in **closed set** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	**0.47**	0.23	**0.45**	0.31	**0.48**	0.36	0.37
	0.63	0.46	**0.65**	**0.48**	**0.63**	0.46	**0.49**
ow	0.33	0.28	0.30	0.33	0.24	0.30	0.19
	0.48	**0.48**	**0.45**	**0.54**	0.42	**0.57**	**0.59**
ay	-0.02	0.09	-0.04	0.05	0.12	0.05	0.08
	0.48	0.22	0.44	0.26	**0.56**	**0.46**	**0.55**
aw	**0.46**	0.22	0.38	0.31	0.40	**0.58**	**0.61**
	0.72	**0.52**	**0.76**	**0.52**	**0.75**	**0.69**	**0.73**
oy	0.40	**0.51**	0.39	**0.56**	0.28	0.39	0.32
	0.12	0.33	0.15	**0.49**	0.32	0.30	**0.75**

Table B.17: Correlation between perceptual scores and weighted distance measures computed using **three** frames of **MCA**s in **open set (9+1)** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.31	0.27	0.40	0.31	0.36	0.32	0.36
	0.54	0.46	**0.58**	**0.46**	**0.58**	**0.47**	**0.48**
ow	-0.00	0.04	0.08	0.15	0.10	0.18	0.07
	0.52	**0.46**	**0.47**	**0.53**	0.42	**0.59**	**0.60**
ay	-0.02	0.06	-0.04	0.04	0.13	0.06	0.09
	0.50	0.21	**0.45**	0.25	**0.56**	**0.45**	**0.56**
aw	0.37	0.28	0.34	0.32	0.37	**0.58**	**0.62**
	0.39	**0.55**	**0.58**	**0.53**	**0.67**	**0.64**	**0.70**
oy	0.32	**0.53**	0.32	**0.58**	0.26	0.40	0.29
	0.12	0.18	0.14	0.40	0.31	0.24	**0.75**

Table B.18: Correlation between perceptual scores and weighted distance measures computed using **three** frames of **MCA**s in **open set (9+1)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.13	0.34	0.15	**0.50**	0.33	0.31	**0.75**
	0.48	0.24	0.45	0.32	**0.51**	0.38	0.38
ow	**0.65**	**0.48**	**0.66**	**0.51**	**0.63**	**0.46**	**0.48**
	0.33	0.28	0.29	0.32	0.23	0.30	0.19
ay	**0.48**	**0.49**	**0.45**	**0.54**	**0.43**	**0.56**	**0.59**
	-0.02	0.09	-0.04	0.05	0.12	0.05	0.08
aw	**0.48**	0.22	0.44	0.25	**0.56**	**0.46**	**0.55**
	0.47	0.23	0.38	0.31	0.41	**0.58**	**0.61**
oy	**0.69**	**0.51**	**0.75**	**0.52**	**0.73**	**0.68**	**0.71**
	0.39	**0.50**	0.38	**0.56**	0.27	0.38	0.31

Table B.19: Correlation between perceptual scores and weighted distance measures computed using **three** frames of **MCA**s in **open set (8+2)** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.15	0.23	0.18	**0.46**	0.36	0.30	**0.76**
	0.47	0.30	**0.47**	0.32	0.45	0.37	0.37
ow	**0.55**	**0.46**	**0.59**	**0.47**	**0.55**	**0.47**	**0.50**
	-0.04	-0.07	0.04	0.00	0.06	0.19	0.08
ay	**0.53**	**0.47**	**0.47**	**0.53**	**0.44**	**0.58**	**0.60**
	-0.02	0.04	-0.04	0.03	0.13	0.06	0.09
aw	**0.50**	0.20	0.45	0.25	**0.56**	**0.45**	**0.56**
	0.35	0.27	0.33	0.32	0.37	**0.57**	**0.62**
oy	0.38	**0.55**	**0.65**	**0.53**	**0.69**	**0.65**	**0.65**
	0.29	**0.52**	0.29	**0.58**	0.26	0.40	0.28

Table B.20: Correlation between perceptual scores and weighted distance measures computed using **three** frames of **MCA**s in **open set (8+2)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.22	0.15	0.35	0.26	0.33	0.30	0.33
	0.43	0.29	0.41	0.33	0.37	0.36	0.27
ow	-0.05	0.06	-0.05	0.04	0.10	0.05	0.04
	0.41	0.15	0.38	0.24	0.36	**0.58**	**0.62**
ay	0.19	**0.42**	0.24	**0.48**	0.10	0.34	0.34
	0.63	**0.47**	**0.65**	**0.49**	**0.63**	**0.47**	**0.49**
aw	**0.47**	**0.49**	**0.45**	**0.54**	0.40	**0.58**	**0.60**
	0.47	0.23	0.44	0.26	**0.55**	**0.46**	**0.56**
oy	**0.74**	**0.53**	**0.77**	**0.53**	**0.76**	**0.70**	**0.73**
	0.12	0.31	0.15	**0.47**	0.31	0.28	**0.75**

Table B.21: Correlation between perceptual scores and weighted distance measures computed using **three** frames of **MCA**s in **open set (5+5)** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	-0.16	0.23	-0.07	0.28	0.09	0.28	0.35
	0.42	0.23	0.41	0.31	0.40	0.35	0.25
ow	-0.07	0.06	-0.07	0.04	0.07	0.04	0.06
	0.33	0.22	0.35	0.27	0.31	**0.57**	**0.62**
ay	-0.11	**0.49**	0.00	**0.51**	-0.10	0.32	0.33
	0.25	**0.52**	0.35	**0.50**	0.34	0.44	**0.47**
aw	0.38	0.43	0.38	**0.51**	0.17	**0.58**	**0.60**
	0.43	0.15	0.38	0.24	0.43	0.42	**0.57**
oy	0.11	**0.57**	0.29	**0.54**	**0.49**	**0.57**	**0.68**
	0.06	0.01	0.08	0.35	0.21	0.16	**0.71**

Table B.22: Correlation between perceptual scores and weighted distance measures computed using **three** frames of **MCA**s in **open set (5+5)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	**0.48**	0.24	**0.45**	0.32	**0.53**	0.38	0.37
	0.64	**0.47**	**0.66**	**0.49**	**0.64**	0.46	**0.49**
ow	0.27	0.20	0.27	0.23	0.28	0.30	0.18
	0.49	**0.48**	**0.46**	**0.54**	0.44	**0.57**	**0.59**
ay	-0.02	0.08	-0.04	0.05	0.13	0.05	0.08
	0.48	0.23	0.44	0.26	**0.57**	**0.46**	**0.55**
aw	**0.46**	0.23	0.38	0.31	0.40	**0.58**	**0.61**
	0.73	**0.52**	**0.77**	**0.52**	**0.76**	**0.70**	**0.74**
oy	0.41	**0.50**	0.39	**0.56**	0.28	0.38	0.35
	0.13	0.34	0.16	**0.50**	0.33	0.31	**0.75**

Table B.23: Correlation between perceptual scores and weighted distance measures computed using **three** frames of **MCA**s in **closed set** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	**0.48**	0.30	**0.49**	0.33	**0.58**	0.43	0.38
	0.56	**0.49**	**0.60**	**0.49**	**0.60**	**0.47**	**0.48**
ow	0.40	0.09	0.36	0.18	0.28	0.25	0.21
	0.53	**0.49**	**0.47**	**0.54**	0.44	**0.59**	**0.60**
ay	-0.02	0.06	-0.04	0.04	0.13	0.07	0.09
	0.50	0.22	**0.45**	0.26	**0.57**	**0.45**	**0.56**
aw	0.41	0.28	0.35	0.33	0.38	**0.58**	**0.62**
	0.60	**0.54**	**0.70**	**0.53**	**0.71**	**0.67**	**0.71**
oy	0.36	**0.51**	0.34	**0.57**	0.31	0.40	0.31
	0.14	0.22	0.17	**0.46**	0.32	0.26	**0.75**

Table B.24: Correlation between perceptual scores and weighted distance measures computed using **three** frames of **MCA**s in **closed set** experiment (with intercept)

B.2 Weighted five frame distances

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	**0.70**	**0.68**	**0.54**	**0.53**	**0.46**	0.41
	0.69	**0.68**	**0.69**	**0.65**	**0.72**	**0.60**
ow	0.20	0.18	0.23	0.16	0.20	0.16
	0.50	**0.52**	**0.53**	**0.56**	**0.56**	**0.51**
ay	0.36	0.36	0.30	0.26	**0.44**	0.30
	0.75	**0.74**	**0.77**	**0.74**	**0.68**	**0.69**
aw	0.07	0.17	0.16	0.28	0.06	0.37
	0.79	**0.77**	**0.80**	**0.73**	**0.81**	**0.75**
oy	-0.15	-0.02	-0.03	0.10	-0.01	0.30
	0.24	0.25	0.19	0.20	0.26	0.27

Table B.25: Correlation between perceptual scores and weighted distance measures computed using **five** frames of **MFCC**s in **open set (9+1)** experiment

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	**0.66**	**0.66**	**0.47**	**0.56**	0.41	0.44
	0.68	**0.67**	**0.67**	**0.59**	**0.69**	**0.53**
ow	0.14	0.13	0.26	0.15	0.18	0.16
	0.37	0.36	**0.47**	0.43	**0.51**	0.30
ay	0.25	0.26	0.19	0.16	**0.43**	0.29
	0.59	**0.60**	**0.61**	**0.62**	**0.61**	**0.65**
aw	-0.11	0.02	0.05	0.19	-0.03	0.35
	0.71	**0.68**	**0.61**	**0.49**	**0.79**	**0.64**
oy	-0.23	-0.09	-0.08	0.03	-0.15	0.24
	0.26	0.29	0.20	0.23	0.25	0.28

Table B.26: Correlation between perceptual scores and weighted distance measures computed using **five** frames of **MFCC**s in **open set (9+1)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	0.22	0.23	0.18	0.20	0.24	0.28
	0.55	**0.56**	**0.51**	**0.51**	0.45	0.45
ow	**0.64**	**0.62**	**0.66**	**0.60**	**0.67**	**0.55**
	0.19	0.18	0.21	0.16	0.20	0.16
ay	**0.50**	**0.53**	**0.53**	**0.56**	**0.56**	**0.49**
	0.35	0.35	0.29	0.26	0.44	0.29
aw	**0.74**	**0.73**	**0.76**	**0.73**	**0.68**	**0.68**
	0.05	0.15	0.16	0.27	0.06	0.37
oy	**0.79**	**0.77**	**0.79**	**0.72**	**0.81**	**0.74**
	-0.16	-0.02	-0.04	0.10	-0.03	0.30

Table B.27: Correlation between perceptual scores and weighted distance measures computed using **five** frames of **MFCC**s in **open set (8+2)** experiment

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	0.21	0.25	0.18	0.23	0.22	0.30
	0.49	**0.55**	**0.47**	**0.56**	0.37	**0.49**
ow	**0.55**	**0.52**	**0.58**	0.45	**0.56**	0.43
	0.12	0.11	0.18	0.12	0.15	0.16
ay	0.37	0.36	**0.47**	**0.45**	**0.52**	0.28
	0.24	0.25	0.19	0.15	0.43	0.26
aw	**0.49**	**0.51**	**0.52**	**0.55**	**0.60**	**0.62**
	-0.17	-0.05	0.00	0.15	-0.05	0.34
oy	**0.69**	**0.68**	**0.59**	**0.48**	**0.78**	**0.62**
	-0.26	-0.12	-0.08	0.02	-0.16	0.20

Table B.28: Correlation between perceptual scores and weighted distance measures computed using **five** frames of **MFCC**s in **open set (8+2)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	**0.74**	**0.69**	**0.50**	0.43	0.44	0.33
	0.21	0.16	0.23	0.08	0.22	0.07
ow	0.35	0.35	0.23	0.14	0.38	0.20
	0.06	0.18	0.16	0.30	0.07	0.40
ay	-0.14	0.03	-0.04	0.11	0.03	0.35
	0.69	**0.67**	**0.69**	**0.63**	**0.68**	**0.59**
aw	**0.52**	**0.55**	**0.51**	**0.55**	**0.56**	**0.60**
	0.72	**0.71**	**0.74**	**0.70**	**0.61**	**0.63**
oy	**0.78**	**0.79**	**0.73**	**0.76**	**0.79**	**0.79**
	0.23	0.25	0.18	0.20	0.24	0.26

Table B.29: Correlation between perceptual scores and weighted distance measures computed using **five** frames of **MFCC**s in **open set (5+5)** experiment

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	**0.73**	**0.67**	0.44	0.35	0.42	0.28
	0.22	0.18	0.31	0.06	0.27	0.04
ow	0.25	0.22	0.09	-0.02	0.31	0.10
	-0.04	0.14	0.10	0.28	0.01	0.41
ay	-0.20	-0.02	-0.09	0.03	-0.07	0.27
	0.65	**0.66**	**0.60**	**0.50**	**0.57**	0.43
aw	**0.49**	**0.50**	**0.47**	**0.45**	**0.53**	0.40
	0.59	**0.58**	**0.60**	**0.46**	**0.46**	**0.50**
oy	**0.73**	**0.76**	**0.60**	**0.64**	**0.75**	**0.73**
	0.25	0.30	0.18	0.22	0.22	0.25

Table B.30: Correlation between perceptual scores and weighted distance measures computed using **five** frames of **MFCC**s in **open set (5+5)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	**0.72**	**0.70**	**0.59**	**0.58**	**0.55**	0.44
	0.69	**0.68**	**0.70**	**0.66**	**0.72**	**0.61**
ow	0.26	0.23	0.29	0.21	0.26	0.18
	0.50	**0.52**	**0.52**	**0.56**	**0.55**	**0.51**
ay	0.37	0.38	0.32	0.30	**0.45**	0.34
	0.77	**0.76**	**0.77**	**0.75**	**0.68**	**0.71**
aw	0.10	0.18	0.17	0.29	0.09	0.38
	0.80	**0.78**	**0.80**	**0.76**	**0.81**	**0.77**
oy	-0.14	-0.01	-0.03	0.11	0.02	0.31
	0.31	0.34	0.22	0.27	0.28	0.32

Table B.31: Correlation between perceptual scores and weighted distance measures computed using **five** frames of **MFCC**s in **closed set** experiment

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	**0.73**	**0.71**	**0.61**	**0.65**	**0.64**	**0.51**
	0.69	**0.68**	**0.68**	**0.63**	**0.71**	**0.58**
ow	0.31	0.26	**0.47**	0.32	0.34	0.21
	0.42	0.42	**0.49**	**0.48**	**0.53**	0.41
ay	0.29	0.31	0.23	0.24	**0.45**	0.35
	0.67	**0.68**	**0.68**	**0.68**	**0.65**	**0.69**
aw	-0.03	0.08	0.11	0.21	0.03	0.37
	0.74	**0.71**	**0.70**	**0.61**	**0.80**	**0.70**
oy	-0.23	-0.07	-0.08	0.05	-0.12	0.27
	0.37	0.42	0.26	0.33	0.28	0.35

Table B.32: Correlation between perceptual scores and weighted distance measures computed using **five** frames of **MFCC**s in **closed set** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.06	0.21	0.14	0.28	0.32	0.17	0.33
	0.67	**0.64**	**0.73**	**0.64**	**0.68**	**0.69**	**0.60**
ow	0.23	0.18	0.23	0.14	0.20	0.34	0.17
	0.47	**0.57**	0.41	**0.52**	0.38	0.44	**0.46**
ay	0.07	0.07	0.05	-0.01	0.36	0.09	0.00
	0.67	**0.67**	**0.61**	**0.67**	**0.77**	**0.56**	**0.58**
aw	0.31	**0.50**	0.21	**0.53**	0.11	0.41	**0.58**
	0.80	**0.79**	**0.80**	**0.78**	**0.81**	**0.77**	**0.73**
oy	0.28	0.36	0.31	0.42	0.29	0.26	0.41
	0.20	0.22	0.20	0.21	0.38	0.25	0.29

Table B.33: Correlation between perceptual scores and weighted distance measures computed using **five** frames of **LSF**s in **open set (9+1)** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.03	0.21	0.12	0.26	0.30	0.15	0.30
	0.62	0.45	**0.66**	0.42	**0.64**	**0.69**	**0.51**
ow	0.16	0.01	0.15	-0.02	0.16	0.35	0.08
	0.47	**0.58**	0.40	**0.51**	0.37	0.42	**0.50**
ay	0.14	0.04	0.06	-0.06	0.37	0.13	-0.06
	0.59	**0.52**	**0.55**	**0.53**	**0.71**	**0.47**	**0.50**
aw	0.19	**0.54**	0.10	**0.62**	0.12	0.25	**0.65**
	0.81	**0.76**	**0.78**	**0.73**	**0.80**	**0.77**	**0.73**
oy	0.17	0.35	0.12	0.38	0.23	0.22	0.41
	0.15	0.21	0.18	0.20	0.36	0.22	0.25

Table B.34: Correlation between perceptual scores and weighted distance measures computed using **five** frames of **LSF**s in **open set (9+1)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.19	0.23	0.19	0.21	0.37	0.24	0.29
	0.05	0.19	0.13	0.28	0.31	0.17	0.31
ow	**0.65**	**0.63**	**0.72**	**0.62**	**0.66**	**0.68**	**0.61**
	0.21	0.17	0.21	0.12	0.19	0.32	0.16
ay	**0.47**	**0.58**	**0.43**	**0.55**	0.38	**0.44**	**0.53**
	0.06	0.07	0.06	0.00	0.35	0.10	0.01
aw	**0.67**	**0.67**	**0.59**	**0.67**	**0.77**	**0.55**	**0.58**
	0.32	**0.58**	0.23	**0.60**	0.12	0.41	**0.66**
oy	**0.79**	**0.79**	**0.80**	**0.78**	**0.81**	**0.77**	**0.72**
	0.28	0.37	0.33	0.44	0.28	0.26	0.44

Table B.35: Correlation between perceptual scores and weighted distance measures computed using **five** frames of **LSF**s in **open set (8+2)** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.14	0.22	0.15	0.20	0.34	0.18	0.25
	0.02	0.19	0.10	0.27	0.28	0.13	0.30
ow	**0.55**	0.41	**0.53**	0.37	**0.59**	**0.68**	**0.52**
	0.09	-0.01	0.07	-0.04	0.13	0.31	0.07
ay	**0.46**	**0.57**	0.40	**0.53**	0.35	**0.43**	**0.59**
	0.15	0.11	0.08	0.02	0.37	0.14	-0.03
aw	**0.46**	**0.50**	**0.49**	**0.56**	**0.72**	0.38	**0.48**
	0.14	**0.50**	0.07	**0.63**	0.14	0.18	**0.67**
oy	**0.79**	**0.78**	**0.77**	**0.75**	**0.79**	**0.77**	**0.72**
	0.09	0.33	0.03	0.35	0.18	0.22	0.38

Table B.36: Correlation between perceptual scores and weighted distance measures computed using **five** frames of **LSF**s in **open set (8+2)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.05	0.17	0.12	0.23	0.30	0.15	0.33
	0.19	0.08	0.25	0.03	0.18	0.37	-0.00
ow	0.03	-0.03	0.03	-0.12	0.33	0.12	-0.08
	0.26	**0.56**	0.20	**0.51**	0.04	0.35	**0.58**
ay	0.31	0.29	0.29	0.26	0.27	0.29	0.28
	0.67	**0.62**	**0.70**	**0.63**	**0.67**	**0.66**	**0.58**
aw	**0.45**	**0.50**	0.38	0.40	0.37	0.43	0.32
	0.66	**0.67**	**0.67**	**0.72**	**0.76**	**0.61**	**0.59**
oy	**0.81**	**0.81**	**0.81**	**0.80**	**0.81**	**0.76**	**0.73**
	0.21	0.23	0.22	0.23	0.39	0.29	0.34

Table B.37: Correlation between perceptual scores and weighted distance measures computed using **five** frames of **LSF**s in **open set (5+5)** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.03	0.11	0.10	0.16	0.27	0.14	0.27
	0.19	-0.07	0.28	-0.13	0.20	0.39	-0.15
ow	0.03	-0.09	0.03	-0.18	0.34	0.16	-0.13
	0.19	**0.50**	0.09	0.44	-0.01	0.25	**0.53**
ay	0.23	0.17	0.21	0.10	0.26	0.30	0.22
	0.61	0.44	**0.65**	**0.46**	**0.62**	**0.67**	**0.51**
aw	0.44	**0.57**	0.36	0.44	0.36	0.43	0.41
	0.62	**0.59**	**0.63**	**0.64**	**0.67**	**0.59**	**0.56**
oy	**0.81**	**0.78**	**0.79**	**0.78**	**0.78**	**0.78**	**0.75**
	0.15	0.21	0.24	0.21	0.37	0.28	0.32

Table B.38: Correlation between perceptual scores and weighted distance measures computed using **five** frames of **LSF**s in **open set (5+5)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.07	0.21	0.16	0.29	0.35	0.18	0.33
	0.69	**0.64**	**0.74**	**0.64**	**0.70**	**0.70**	**0.60**
ow	0.24	0.21	0.23	0.15	0.25	0.33	0.19
	0.46	**0.55**	0.41	**0.51**	0.39	0.44	**0.47**
ay	0.08	0.09	0.07	0.04	0.36	0.09	0.09
	0.69	**0.71**	**0.64**	**0.72**	**0.77**	**0.59**	**0.66**
aw	0.33	**0.54**	0.22	**0.55**	0.15	0.44	**0.59**
	0.81	**0.81**	**0.80**	**0.80**	**0.81**	**0.78**	**0.72**
oy	0.29	0.37	0.30	0.43	0.29	0.26	0.42
	0.21	0.22	0.21	0.24	0.38	0.27	0.30

Table B.39: Correlation between perceptual scores and weighted distance measures computed using **five** frames of **LSF**s in **closed set** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.05	0.24	0.17	0.28	0.36	0.16	0.30
	0.68	**0.49**	**0.73**	**0.48**	**0.68**	**0.71**	**0.50**
ow	0.35	0.11	0.27	0.04	0.31	0.38	0.14
	0.49	**0.64**	0.43	**0.58**	0.40	0.43	**0.61**
ay	0.15	0.05	0.10	-0.01	0.38	0.14	0.02
	0.65	**0.59**	**0.60**	**0.60**	**0.73**	**0.53**	**0.58**
aw	0.25	**0.63**	0.20	**0.63**	0.22	0.34	**0.65**
	0.82	**0.79**	**0.79**	**0.76**	**0.80**	**0.78**	**0.74**
oy	0.22	0.35	0.23	0.42	0.26	0.27	0.43
	0.15	0.21	0.21	0.25	0.37	0.27	0.28

Table B.40: Correlation between perceptual scores and weighted distance measures computed using **five** frames of **LSF**s in **closed set** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	**0.45**	0.25	0.44	0.32	**0.46**	0.30	0.35
	0.69	**0.51**	**0.74**	**0.51**	**0.67**	**0.53**	**0.51**
ow	0.35	0.22	0.35	0.30	0.31	0.29	0.17
	0.58	**0.51**	**0.52**	**0.55**	**0.51**	**0.60**	**0.62**
ay	-0.05	-0.03	-0.07	-0.02	0.11	0.01	0.01
	0.52	0.24	**0.48**	0.26	**0.60**	**0.50**	**0.54**
aw	0.38	0.20	0.29	0.26	0.21	**0.51**	**0.55**
	0.73	**0.54**	**0.76**	**0.54**	**0.76**	**0.71**	**0.73**
oy	0.41	**0.54**	0.36	**0.57**	0.28	0.46	0.32
	0.16	0.37	0.16	**0.47**	0.34	0.32	**0.73**

Table B.41: Correlation between perceptual scores and weighted distance measures computed using **five** frames of **MCA**s in **open set (9+1)** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.21	0.25	0.23	0.29	0.39	0.30	0.33
	0.52	**0.52**	**0.64**	**0.50**	**0.59**	**0.48**	**0.49**
ow	0.13	0.03	0.17	0.18	0.26	0.28	0.03
	0.66	**0.56**	**0.61**	**0.57**	**0.58**	**0.63**	**0.62**
ay	-0.12	0.10	-0.13	0.08	0.06	-0.00	-0.01
	0.35	0.18	0.38	0.19	**0.45**	0.40	**0.53**
aw	-0.01	0.24	0.06	0.31	0.00	0.34	0.40
	0.50	**0.54**	**0.55**	**0.50**	**0.65**	**0.69**	**0.70**
oy	0.14	**0.59**	0.04	**0.60**	0.15	0.38	0.27
	0.16	-0.01	0.16	0.09	0.33	0.27	**0.67**

Table B.42: Correlation between perceptual scores and weighted distance measures computed using **five** frames of **MCA**s in **open set (9+1)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.16	0.38	0.16	**0.48**	0.35	0.33	**0.74**
	0.46	0.26	0.44	0.32	**0.48**	0.31	0.35
ow	**0.68**	**0.52**	**0.73**	**0.52**	**0.67**	**0.53**	**0.50**
	0.31	0.23	0.31	0.31	0.28	0.27	0.13
ay	**0.58**	**0.51**	**0.53**	**0.55**	**0.52**	**0.61**	**0.63**
	-0.06	-0.03	-0.07	-0.01	0.11	0.00	0.01
aw	**0.52**	0.23	**0.48**	0.26	**0.61**	**0.50**	**0.54**
	0.39	0.20	0.30	0.27	0.23	**0.52**	**0.55**
oy	**0.70**	**0.53**	**0.74**	**0.53**	**0.74**	**0.71**	**0.72**
	0.41	**0.53**	0.35	**0.57**	0.27	**0.45**	0.32

Table B.43: Correlation between perceptual scores and weighted distance measures computed using **five** frames of **MCA**s in **open set (8+2)** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.18	-0.01	0.18	0.07	0.36	0.31	**0.61**
	0.38	0.26	0.35	0.28	0.46	0.33	0.31
ow	**0.55**	**0.54**	**0.59**	**0.49**	**0.58**	**0.51**	**0.47**
	0.06	-0.05	0.07	0.05	0.18	0.25	-0.01
ay	**0.66**	**0.56**	**0.62**	**0.58**	**0.58**	**0.63**	**0.62**
	-0.12	0.08	-0.13	0.07	0.07	-0.00	-0.01
aw	0.25	0.19	0.36	0.19	**0.48**	0.35	**0.48**
	-0.08	0.24	0.03	0.31	0.02	0.30	0.34
oy	**0.50**	**0.54**	**0.55**	**0.49**	**0.66**	**0.69**	**0.67**
	0.09	**0.56**	-0.04	**0.58**	0.13	0.37	0.26

Table B.44: Correlation between perceptual scores and weighted distance measures computed using **five** frames of **MCA**s in **open set (8+2)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.27	0.19	0.36	0.27	0.38	0.28	0.31
	0.46	0.23	**0.48**	0.31	0.46	0.39	0.37
ow	-0.06	-0.08	-0.07	-0.04	0.08	0.03	0.01
	0.42	0.14	0.36	0.21	0.32	**0.54**	**0.60**
ay	0.16	**0.46**	0.20	**0.48**	0.09	0.40	0.29
	0.69	**0.51**	**0.73**	**0.52**	**0.67**	**0.54**	**0.50**
aw	**0.57**	**0.52**	**0.52**	**0.56**	**0.49**	**0.61**	**0.62**
	0.52	0.24	**0.50**	0.26	**0.62**	**0.50**	**0.56**
oy	**0.74**	**0.54**	**0.76**	**0.54**	**0.77**	**0.71**	**0.70**
	0.16	0.34	0.16	0.44	0.34	0.32	**0.72**

Table B.45: Correlation between perceptual scores and weighted distance measures computed using **five** frames of **MCA**s in **open set (5+5)** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	-0.07	0.28	-0.04	0.29	0.18	0.27	0.32
	0.56	0.11	**0.54**	0.30	**0.54**	**0.50**	0.40
ow	-0.12	-0.11	-0.11	-0.02	0.03	-0.00	0.02
	0.25	0.19	0.33	0.24	0.28	**0.48**	**0.59**
ay	-0.10	**0.57**	-0.11	**0.54**	-0.14	0.35	0.23
	0.38	0.44	**0.49**	**0.45**	0.42	0.40	0.38
aw	**0.66**	**0.49**	**0.59**	**0.50**	**0.50**	**0.63**	**0.58**
	0.33	0.04	0.36	0.02	0.36	0.36	**0.49**
oy	0.40	0.37	0.41	0.39	**0.52**	**0.65**	**0.63**
	0.12	-0.24	0.12	-0.13	0.27	0.19	**0.63**

Table B.46: Correlation between perceptual scores and weighted distance measures computed using **five** frames of **MCA**s in **open set (5+5)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	**0.45**	0.26	0.44	0.32	**0.50**	0.32	0.35
	0.69	**0.51**	**0.74**	**0.51**	**0.68**	**0.53**	**0.51**
ow	0.38	0.17	0.40	0.23	0.39	0.36	0.28
	0.57	**0.51**	**0.52**	**0.55**	**0.52**	**0.60**	**0.62**
ay	-0.05	-0.03	-0.07	-0.02	0.11	0.02	0.02
	0.52	0.24	**0.49**	0.27	**0.62**	**0.50**	**0.55**
aw	0.38	0.21	0.29	0.28	0.24	**0.50**	**0.55**
	0.74	**0.54**	**0.76**	**0.54**	**0.77**	**0.72**	**0.73**
oy	0.42	**0.54**	0.37	**0.58**	0.29	0.46	0.39
	0.16	0.38	0.17	**0.48**	0.35	0.33	**0.73**

Table B.47: Correlation between perceptual scores and weighted distance measures computed using **five** frames of **MCA**s in **closed set** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.35	0.30	0.37	0.32	**0.54**	0.41	0.34
	0.52	**0.55**	**0.64**	**0.52**	**0.60**	**0.47**	**0.48**
ow	**0.54**	0.17	**0.57**	0.27	**0.50**	**0.46**	0.37
	0.66	**0.58**	**0.62**	**0.58**	**0.59**	**0.63**	**0.62**
ay	-0.12	0.18	-0.13	0.14	0.07	0.01	0.01
	0.42	0.21	0.42	0.20	**0.51**	0.43	**0.54**
aw	0.19	0.27	0.21	0.34	0.16	0.44	**0.52**
	0.62	**0.55**	**0.63**	**0.52**	**0.70**	**0.71**	**0.71**
oy	0.22	**0.59**	0.11	**0.61**	0.21	0.40	0.35
	0.18	0.08	0.17	0.22	0.34	0.28	**0.72**

Table B.48: Correlation between perceptual scores and weighted distance measures computed using **five** frames of **MCA**s in **closed set** experiment (with intercept)

B.3 Weighted seven frame distances

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	**0.51**	**0.56**	0.42	**0.52**	0.38	0.44
	0.71	**0.70**	**0.71**	**0.64**	**0.73**	**0.61**
ow	0.18	0.15	0.19	0.12	0.21	0.11
	0.55	**0.56**	**0.58**	**0.59**	**0.57**	**0.51**
ay	0.35	0.36	0.28	0.24	**0.44**	0.23
	0.78	**0.76**	**0.77**	**0.75**	**0.68**	**0.71**
aw	0.02	0.10	0.13	0.25	0.08	0.37
	0.83	**0.80**	**0.82**	**0.72**	**0.82**	**0.71**
oy	0.09	0.12	0.01	0.09	0.04	0.28
	0.23	0.23	0.18	0.19	0.25	0.28

Table B.49: Correlation between perceptual scores and weighted distance measures computed using **seven** frames of **MFCC**s in **open set (9+1)** experiment

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	0.34	**0.48**	0.23	**0.51**	0.19	0.44
	0.70	**0.67**	**0.70**	**0.59**	**0.71**	**0.50**
ow	0.12	0.11	0.20	0.12	0.19	0.12
	0.42	0.35	**0.49**	0.38	**0.49**	0.20
ay	0.25	0.28	0.17	0.16	**0.44**	0.26
	0.60	**0.55**	**0.64**	**0.60**	**0.58**	**0.61**
aw	-0.13	-0.03	0.01	0.17	-0.01	0.37
	0.78	**0.72**	**0.65**	0.45	**0.83**	**0.61**
oy	-0.06	-0.07	-0.07	-0.05	-0.05	0.09
	0.23	0.26	0.18	0.21	0.24	0.30

Table B.50: Correlation between perceptual scores and weighted distance measures computed using **seven** frames of **MFCC**s in **open set (9+1)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	0.21	0.22	0.17	0.19	0.23	0.29
	0.48	**0.52**	0.44	**0.48**	0.36	0.45
ow	**0.67**	**0.64**	**0.67**	**0.58**	**0.69**	**0.57**
	0.17	0.15	0.18	0.13	0.21	0.11
ay	**0.56**	**0.57**	**0.58**	**0.58**	**0.57**	**0.46**
	0.35	0.36	0.27	0.23	**0.45**	0.21
aw	**0.77**	**0.74**	**0.76**	**0.74**	**0.69**	**0.70**
	0.01	0.09	0.14	0.24	0.08	0.36
oy	**0.83**	**0.79**	**0.81**	**0.70**	**0.82**	**0.70**
	0.05	0.10	-0.02	0.07	0.01	0.29

Table B.51: Correlation between perceptual scores and weighted distance measures computed using **seven** frames of **MFCC**s in **open set (8+2)** experiment

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	0.18	0.21	0.16	0.19	0.19	0.31
	0.33	0.43	0.28	0.46	0.14	0.45
ow	**0.59**	**0.55**	**0.59**	0.45	**0.59**	**0.45**
	0.11	0.10	0.14	0.10	0.16	0.13
ay	**0.42**	0.33	**0.49**	0.38	**0.49**	0.15
	0.25	0.27	0.17	0.15	0.44	0.22
aw	**0.50**	**0.46**	**0.56**	**0.54**	**0.55**	**0.58**
	-0.18	-0.08	-0.04	0.14	-0.03	0.34
oy	**0.73**	**0.67**	**0.56**	0.35	**0.80**	**0.58**
	-0.16	-0.16	-0.11	-0.11	-0.05	0.03

Table B.52: Correlation between perceptual scores and weighted distance measures computed using **seven** frames of **MFCC**s in **open set (8+2)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	0.36	**0.45**	0.32	0.39	0.34	0.37
	0.17	0.12	0.18	0.06	0.22	0.03
ow	0.34	0.35	0.22	0.13	0.37	0.15
	-0.02	0.10	0.11	0.27	0.08	0.40
ay	0.25	0.30	0.06	0.14	0.08	0.36
	0.71	**0.68**	**0.72**	**0.64**	**0.70**	**0.59**
aw	**0.53**	**0.56**	**0.53**	**0.56**	**0.56**	**0.59**
	0.73	**0.71**	**0.76**	**0.71**	**0.61**	**0.63**
oy	**0.80**	**0.81**	**0.75**	**0.78**	**0.81**	**0.79**
	0.22	0.23	0.17	0.19	0.24	0.26

Table B.53: Correlation between perceptual scores and weighted distance measures computed using **seven** frames of **MFCC**s in **open set (5+5)** experiment

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	0.26	0.39	0.13	0.30	0.23	0.35
	0.16	0.12	0.18	0.04	0.31	0.01
ow	0.24	0.24	0.07	-0.03	0.31	0.07
	-0.09	0.04	0.01	0.26	0.04	**0.45**
ay	0.25	0.25	0.12	0.06	0.04	0.22
	0.67	**0.64**	**0.61**	**0.50**	**0.59**	0.36
aw	**0.46**	**0.45**	0.44	0.41	**0.50**	0.34
	0.55	**0.50**	**0.58**	0.42	0.45	0.43
oy	**0.78**	**0.78**	**0.62**	**0.67**	**0.79**	**0.76**
	0.21	0.26	0.17	0.21	0.23	0.28

Table B.54: Correlation between perceptual scores and weighted distance measures computed using **seven** frames of **MFCC**s in **open set (5+5)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	**0.55**	**0.59**	**0.48**	**0.56**	**0.49**	**0.48**
	0.72	**0.70**	**0.72**	**0.64**	**0.73**	**0.61**
ow	0.24	0.21	0.27	0.20	0.27	0.16
	0.56	**0.56**	**0.58**	**0.59**	**0.57**	**0.52**
ay	0.37	0.38	0.30	0.28	**0.45**	0.29
	0.79	**0.77**	**0.78**	**0.75**	**0.69**	**0.73**
aw	0.05	0.13	0.15	0.26	0.10	0.39
	0.84	**0.81**	**0.82**	**0.75**	**0.83**	**0.74**
oy	0.10	0.15	0.03	0.12	0.07	0.31
	0.30	0.32	0.23	0.27	0.27	0.32

Table B.55: Correlation between perceptual scores and weighted distance measures computed using **seven** frames of **MFCC**s in **closed set** experiment

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	**0.56**	**0.63**	**0.47**	**0.66**	**0.54**	**0.53**
	0.72	**0.69**	**0.72**	**0.63**	**0.74**	**0.55**
ow	0.29	0.26	0.44	0.33	0.36	0.24
	0.48	0.43	**0.53**	**0.46**	**0.53**	0.35
ay	0.30	0.34	0.21	0.24	**0.45**	0.36
	0.69	**0.66**	**0.71**	**0.67**	**0.64**	**0.68**
aw	-0.07	0.04	0.08	0.22	0.05	0.41
	0.81	**0.75**	**0.75**	**0.60**	**0.84**	**0.69**
oy	0.07	0.06	0.02	0.04	0.01	0.19
	0.33	0.38	0.25	0.33	0.27	0.38

Table B.56: Correlation between perceptual scores and weighted distance measures computed using **seven** frames of **MFCC**s in **closed set** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.09	0.28	0.14	0.34	0.34	0.18	0.40
	0.70	**0.64**	**0.73**	**0.62**	**0.71**	**0.73**	**0.60**
ow	0.21	0.11	0.20	0.14	0.20	0.34	0.16
	0.51	**0.54**	**0.47**	**0.49**	0.39	**0.50**	0.42
ay	-0.04	-0.09	-0.03	-0.10	0.32	0.03	-0.12
	0.59	**0.69**	**0.53**	**0.65**	**0.69**	**0.53**	**0.58**
aw	0.32	**0.53**	0.24	**0.55**	0.16	0.40	**0.59**
	0.83	**0.83**	**0.82**	**0.83**	**0.82**	**0.81**	**0.74**
oy	0.33	0.33	0.32	0.40	0.32	0.33	**0.49**
	0.18	0.22	0.19	0.22	0.36	0.23	0.29

Table B.57: Correlation between perceptual scores and weighted distance measures computed using **seven** frames of **LSF**s in **open set (9+1)** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.02	0.23	0.11	0.30	0.29	0.09	0.35
	0.69	**0.48**	**0.71**	0.44	**0.70**	**0.73**	**0.49**
ow	0.18	-0.03	0.12	0.02	0.16	0.35	0.05
	0.53	**0.57**	**0.49**	**0.51**	0.39	**0.48**	0.44
ay	-0.07	-0.15	-0.06	-0.16	0.32	0.02	-0.20
	0.45	**0.53**	**0.48**	**0.50**	**0.57**	**0.45**	**0.50**
aw	0.29	**0.58**	0.17	**0.65**	0.18	0.29	**0.69**
	0.80	**0.76**	**0.72**	**0.74**	**0.80**	**0.79**	**0.74**
oy	0.35	0.35	0.18	0.41	0.33	0.32	**0.50**
	0.11	0.21	0.19	0.21	0.33	0.22	0.26

Table B.58: Correlation between perceptual scores and weighted distance measures computed using **seven** frames of **LSF**s in **open set (9+1)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.19	0.23	0.18	0.22	0.35	0.23	0.29
	0.07	0.24	0.14	0.32	0.33	0.20	0.37
ow	**0.67**	**0.63**	**0.71**	**0.59**	**0.70**	**0.72**	**0.60**
	0.19	0.10	0.18	0.14	0.18	0.33	0.16
ay	**0.51**	**0.53**	**0.48**	**0.50**	0.39	**0.50**	**0.46**
	-0.04	-0.09	-0.03	-0.10	0.32	0.04	-0.13
aw	**0.56**	**0.70**	**0.50**	**0.66**	**0.70**	**0.50**	**0.59**
	0.32	**0.58**	0.26	**0.60**	0.20	0.39	**0.67**
oy	**0.83**	**0.83**	**0.82**	**0.84**	**0.82**	**0.81**	**0.75**
	0.33	0.35	0.33	0.43	0.31	0.32	**0.51**

Table B.59: Correlation between perceptual scores and weighted distance measures computed using **seven** frames of **LSF**s in **open set (8+2)** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.11	0.22	0.16	0.21	0.32	0.19	0.26
	0.01	0.21	0.10	0.30	0.27	0.11	0.34
ow	**0.64**	**0.46**	**0.65**	0.40	**0.65**	**0.73**	**0.52**
	0.11	-0.03	0.07	0.02	0.13	0.31	0.07
ay	**0.52**	**0.53**	**0.49**	**0.50**	0.38	**0.50**	**0.47**
	-0.07	-0.11	-0.05	-0.12	0.32	0.05	-0.19
aw	0.41	**0.52**	**0.50**	**0.53**	**0.65**	0.32	**0.49**
	0.29	**0.56**	0.18	**0.63**	0.28	0.21	**0.72**
oy	**0.77**	**0.77**	**0.70**	**0.75**	**0.79**	**0.78**	**0.74**
	0.30	0.34	0.14	0.41	0.30	0.31	**0.48**

Table B.60: Correlation between perceptual scores and weighted distance measures computed using **seven** frames of **LSF**s in **open set (8+2)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.11	0.29	0.12	0.32	0.33	0.17	0.43
	0.11	-0.06	0.19	0.02	0.17	0.37	-0.04
ow	-0.08	-0.20	-0.06	-0.20	0.32	0.08	-0.18
	0.27	**0.61**	0.23	**0.57**	0.05	0.36	**0.66**
ay	0.26	0.17	0.29	0.20	0.23	0.35	0.31
	0.71	**0.63**	**0.73**	**0.64**	**0.72**	**0.70**	**0.57**
aw	**0.50**	**0.50**	0.43	0.42	0.41	**0.49**	0.36
	0.61	**0.69**	**0.64**	**0.72**	**0.70**	**0.59**	**0.58**
oy	**0.81**	**0.82**	**0.80**	**0.81**	**0.81**	**0.79**	**0.73**
	0.18	0.23	0.23	0.24	0.36	0.28	0.35

Table B.61: Correlation between perceptual scores and weighted distance measures computed using **seven** frames of **LSF**s in **open set (5+5)** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.03	0.23	0.07	0.22	0.23	0.09	0.34
	0.21	-0.13	0.21	-0.12	0.20	0.38	-0.18
ow	-0.04	-0.18	-0.05	-0.21	0.35	0.14	-0.18
	0.25	**0.54**	0.16	**0.51**	-0.04	0.30	**0.55**
ay	0.40	0.10	0.29	0.08	0.24	0.37	0.11
	0.67	0.41	**0.69**	**0.46**	**0.71**	**0.71**	**0.48**
aw	**0.52**	**0.61**	0.44	**0.49**	0.42	**0.48**	**0.47**
	0.57	**0.57**	**0.60**	**0.60**	**0.55**	**0.57**	**0.55**
oy	**0.80**	**0.75**	**0.75**	**0.75**	**0.79**	**0.82**	**0.73**
	0.11	0.19	0.26	0.22	0.31	0.26	0.33

Table B.62: Correlation between perceptual scores and weighted distance measures computed using **seven** frames of **LSF**s in **open set (5+5)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.11	0.29	0.15	0.35	0.37	0.20	0.40
	0.71	**0.65**	**0.74**	**0.62**	**0.71**	**0.73**	**0.61**
ow	0.22	0.14	0.21	0.16	0.24	0.34	0.18
	0.49	**0.52**	**0.45**	**0.49**	0.40	**0.49**	0.45
ay	-0.00	-0.05	-0.01	-0.07	0.33	0.04	-0.04
	0.65	**0.71**	**0.58**	**0.68**	**0.72**	**0.57**	**0.65**
aw	0.36	**0.58**	0.27	**0.57**	0.22	0.44	**0.61**
	0.84	**0.83**	**0.82**	**0.84**	**0.82**	**0.81**	**0.74**
oy	0.36	0.37	0.31	0.45	0.34	0.33	**0.54**
	0.18	0.23	0.20	0.24	0.36	0.26	0.31

Table B.63: Correlation between perceptual scores and weighted distance measures computed using **seven** frames of **LSF**s in **closed set** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.06	0.29	0.14	0.33	0.37	0.13	0.37
	0.71	**0.52**	**0.73**	**0.48**	**0.73**	**0.75**	**0.53**
ow	0.35	0.08	0.22	0.09	0.30	0.37	0.12
	0.54	**0.63**	**0.51**	**0.58**	0.43	**0.49**	**0.57**
ay	0.03	-0.10	0.01	-0.12	0.34	0.09	-0.12
	0.57	**0.60**	**0.54**	**0.57**	**0.68**	**0.52**	**0.59**
aw	0.35	**0.70**	0.26	**0.69**	0.32	0.39	**0.70**
	0.82	**0.80**	**0.76**	**0.79**	**0.81**	**0.80**	**0.75**
oy	0.40	0.38	0.28	0.46	0.37	0.38	**0.53**
	0.10	0.22	0.22	0.26	0.35	0.27	0.31

Table B.64: Correlation between perceptual scores and weighted distance measures computed using **seven** frames of **LSF**s in **closed set** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.44	0.24	0.43	0.32	**0.45**	0.25	0.36
	0.66	**0.48**	**0.73**	**0.49**	**0.65**	**0.56**	**0.53**
ow	0.37	0.25	0.32	0.29	0.32	0.30	0.17
	0.57	**0.61**	**0.53**	**0.62**	**0.52**	**0.60**	**0.64**
ay	-0.03	-0.01	-0.07	-0.01	0.11	0.02	-0.02
	0.52	0.21	**0.50**	0.25	**0.61**	**0.50**	**0.57**
aw	0.35	0.24	0.29	0.31	0.19	**0.48**	**0.53**
	0.70	**0.53**	**0.73**	**0.54**	**0.73**	**0.68**	**0.74**
oy	**0.50**	0.44	0.42	**0.51**	0.29	**0.54**	0.32
	0.30	**0.50**	0.29	**0.58**	0.41	**0.46**	**0.73**

Table B.65: Correlation between perceptual scores and weighted distance measures computed using **seven** frames of **MCA**s in **open set (9+1)** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.15	0.27	0.16	0.30	0.38	0.27	0.33
	0.50	0.45	**0.65**	0.45	**0.58**	**0.51**	0.36
ow	0.05	0.05	0.04	0.07	0.26	0.23	0.04
	0.64	**0.66**	**0.62**	**0.63**	**0.60**	**0.65**	**0.67**
ay	-0.11	0.14	-0.14	0.09	0.04	-0.04	-0.05
	0.36	0.18	0.41	0.20	**0.47**	0.43	**0.50**
aw	-0.05	0.26	0.05	0.32	-0.00	0.35	**0.47**
	0.49	**0.55**	**0.52**	**0.52**	**0.61**	**0.71**	**0.69**
oy	0.17	0.21	0.02	0.35	0.11	0.33	0.37
	0.12	-0.11	0.09	-0.01	0.27	0.15	**0.64**

Table B.66: Correlation between perceptual scores and weighted distance measures computed using **seven** frames of **MCA**s in **open set (9+1)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.30	**0.52**	0.28	**0.60**	0.39	**0.47**	**0.74**
	0.44	0.23	0.43	0.31	**0.48**	0.26	0.36
ow	**0.68**	**0.49**	**0.70**	**0.49**	**0.64**	**0.56**	**0.49**
	0.33	0.26	0.27	0.28	0.28	0.27	0.12
ay	**0.57**	**0.63**	**0.54**	**0.63**	**0.54**	**0.61**	**0.64**
	-0.04	-0.01	-0.07	-0.01	0.11	0.00	-0.03
aw	**0.52**	0.20	**0.50**	0.23	**0.61**	**0.49**	**0.57**
	0.36	0.25	0.30	0.32	0.21	**0.48**	**0.53**
oy	**0.69**	**0.53**	**0.72**	**0.53**	**0.72**	**0.68**	**0.72**
	0.49	0.43	0.41	**0.51**	0.30	**0.54**	0.29

Table B.67: Correlation between perceptual scores and weighted distance measures computed using **seven** frames of **MCA**s in **open set (8+2)** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.18	-0.04	0.15	0.01	0.29	0.22	**0.54**
	0.38	0.32	0.35	0.31	**0.47**	0.35	0.31
ow	**0.50**	0.44	0.35	0.40	**0.54**	**0.54**	0.39
	-0.01	-0.01	-0.04	-0.08	0.17	0.21	-0.02
ay	**0.65**	**0.67**	**0.64**	**0.65**	**0.61**	**0.65**	**0.67**
	-0.10	0.10	-0.13	0.08	0.05	-0.04	-0.04
aw	0.27	0.19	0.38	0.20	**0.49**	0.36	0.44
	-0.13	0.27	0.02	0.33	0.01	0.29	0.40
oy	**0.48**	**0.54**	**0.51**	**0.50**	**0.61**	**0.72**	**0.68**
	0.11	0.15	-0.05	0.32	0.09	0.30	0.38

Table B.68: Correlation between perceptual scores and weighted distance measures computed using **seven** frames of **MCA**s in **open set (8+2)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.10	0.18	0.28	0.27	0.35	0.27	0.32
	0.42	0.25	**0.47**	0.31	0.45	0.39	0.37
ow	-0.02	-0.06	-0.05	-0.03	0.09	0.03	-0.00
	0.34	0.16	0.32	0.23	0.31	**0.53**	**0.62**
ay	0.20	**0.43**	0.22	**0.46**	0.09	**0.46**	0.35
	0.67	**0.50**	**0.73**	**0.51**	**0.64**	**0.58**	**0.54**
aw	**0.55**	**0.61**	**0.52**	**0.62**	**0.52**	**0.59**	**0.63**
	0.52	0.22	**0.52**	0.24	**0.63**	**0.50**	**0.59**
oy	**0.72**	**0.54**	**0.73**	**0.54**	**0.74**	**0.68**	**0.71**
	0.34	**0.49**	0.35	**0.56**	**0.47**	**0.50**	**0.72**

Table B.69: Correlation between perceptual scores and weighted distance measures computed using **seven** frames of **MCA**s in **open set (5+5)** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	-0.03	0.29	0.06	0.31	0.21	0.31	0.35
	0.56	0.19	**0.49**	0.19	**0.56**	0.39	0.36
ow	-0.12	-0.05	-0.12	0.02	0.03	-0.06	-0.03
	0.27	0.22	0.41	0.30	0.28	**0.49**	**0.62**
ay	-0.09	0.22	-0.12	0.24	-0.17	0.36	0.33
	0.32	0.36	0.43	0.38	0.43	0.35	0.18
aw	**0.60**	**0.49**	**0.54**	0.37	**0.50**	**0.63**	**0.66**
	0.29	0.23	0.28	0.14	0.36	0.33	0.38
oy	0.42	0.28	0.41	0.31	**0.53**	**0.66**	**0.61**
	-0.06	-0.37	-0.12	-0.28	0.22	-0.02	**0.58**

Table B.70: Correlation between perceptual scores and weighted distance measures computed using **seven** frames of **MCA**s in **open set (5+5)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.45	0.25	0.44	0.32	**0.50**	0.28	0.36
	0.67	**0.50**	**0.74**	**0.50**	**0.66**	**0.56**	**0.52**
ow	0.38	0.19	0.40	0.20	0.40	0.35	0.28
	0.56	**0.59**	**0.53**	**0.61**	**0.53**	**0.59**	**0.64**
ay	-0.03	-0.02	-0.06	-0.01	0.12	0.02	-0.01
	0.52	0.23	**0.51**	0.26	**0.63**	**0.51**	**0.57**
aw	0.36	0.26	0.30	0.33	0.22	**0.47**	**0.55**
	0.72	**0.53**	**0.74**	**0.54**	**0.76**	**0.69**	**0.74**
oy	**0.51**	0.44	0.44	**0.51**	0.33	**0.55**	0.40
	0.30	**0.50**	0.31	**0.58**	0.42	**0.47**	**0.72**

Table B.71: Correlation between perceptual scores and weighted distance measures computed using **seven** frames of **MCA**s in **closed set** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.32	0.35	0.34	0.35	**0.52**	0.39	0.35
	0.51	**0.50**	**0.66**	**0.48**	**0.61**	**0.50**	0.36
ow	**0.53**	0.19	**0.54**	0.22	**0.50**	0.44	0.43
	0.65	**0.72**	**0.64**	**0.67**	**0.62**	**0.65**	**0.67**
ay	-0.11	0.22	-0.13	0.15	0.06	-0.02	-0.02
	0.42	0.23	**0.45**	0.24	**0.54**	**0.47**	**0.52**
aw	0.17	0.29	0.24	0.36	0.18	**0.47**	**0.60**
	0.61	**0.56**	**0.61**	**0.53**	**0.67**	**0.72**	**0.73**
oy	0.28	0.40	0.11	**0.49**	0.20	0.41	0.46
	0.34	0.11	0.32	0.23	0.38	0.31	**0.68**

Table B.72: Correlation between perceptual scores and weighted distance measures computed using **seven** frames of **MCA**s in **closed set** experiment (with intercept)

B.4 Weighted Nine frame distances

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	**0.51**	**0.50**	0.37	0.40	0.42	0.38
	0.76	**0.70**	**0.75**	**0.61**	**0.73**	**0.58**
ow	0.20	0.21	0.27	0.26	0.22	0.21
	0.38	0.34	0.34	0.27	0.41	0.33
ay	0.38	0.35	0.34	0.25	**0.43**	0.24
	0.77	**0.75**	**0.71**	**0.72**	**0.70**	**0.75**
aw	0.06	0.10	0.10	0.06	0.11	0.21
	0.82	**0.80**	**0.78**	**0.71**	**0.83**	**0.75**
oy	0.32	0.21	0.32	0.16	0.26	0.19
	0.21	0.23	0.22	0.23	0.25	0.29

Table B.73: Correlation between perceptual scores and weighted distance measures computed using **nine** frames of **MFCC**s in **open set (9+1)** experiment

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	0.31	0.39	0.18	0.36	0.17	0.33
	0.74	**0.68**	**0.76**	**0.61**	**0.71**	**0.49**
ow	0.16	0.17	0.28	0.26	0.20	0.20
	0.24	0.17	0.23	0.12	0.25	0.11
ay	0.28	0.28	0.24	0.18	**0.42**	0.26
	0.63	**0.59**	**0.63**	**0.67**	**0.63**	**0.64**
aw	-0.12	-0.03	-0.02	-0.02	0.02	0.21
	0.78	**0.73**	**0.60**	**0.49**	**0.83**	**0.65**
oy	0.05	-0.02	0.14	-0.01	0.09	0.04
	0.21	0.25	0.21	0.24	0.23	0.31

Table B.74: Correlation between perceptual scores and weighted distance measures computed using **nine** frames of **MFCC**s in **open set (9+1)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	0.20	0.22	0.21	0.23	0.23	0.30
	0.50	**0.47**	0.42	0.37	0.38	0.38
ow	**0.70**	**0.66**	**0.70**	**0.57**	**0.67**	**0.56**
	0.19	0.20	0.26	0.24	0.21	0.20
ay	0.35	0.28	0.32	0.23	0.36	0.29
	0.37	0.34	0.35	0.27	0.42	0.24
aw	**0.77**	**0.75**	**0.70**	**0.71**	**0.71**	**0.73**
	0.06	0.09	0.10	0.04	0.11	0.20
oy	**0.81**	**0.78**	**0.76**	**0.69**	**0.82**	**0.75**
	0.28	0.18	0.30	0.12	0.22	0.20

Table B.75: Correlation between perceptual scores and weighted distance measures computed using **nine** frames of **MFCC**s in **open set (8+2)** experiment

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	0.17	0.21	0.20	0.22	0.19	0.32
	0.32	0.38	0.25	0.32	0.14	0.34
ow	**0.64**	**0.60**	**0.65**	**0.51**	**0.59**	**0.47**
	0.14	0.14	0.22	0.19	0.17	0.17
ay	0.18	0.10	0.22	0.09	0.18	0.06
	0.27	0.26	0.27	0.20	0.38	0.24
aw	**0.52**	**0.50**	**0.59**	**0.60**	**0.61**	**0.65**
	-0.17	-0.08	-0.06	-0.06	0.00	0.20
oy	**0.73**	**0.67**	**0.54**	0.39	**0.80**	**0.61**
	-0.03	-0.10	0.10	-0.07	0.08	-0.01

Table B.76: Correlation between perceptual scores and weighted distance measures computed using **nine** frames of **MFCC**s in **open set (8+2)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	0.42	0.41	0.29	0.32	0.35	0.32
	0.17	0.20	0.26	0.23	0.22	0.16
ow	0.39	0.36	0.27	0.20	0.35	0.18
	0.01	0.07	0.05	0.04	0.10	0.21
ay	**0.49**	0.38	**0.42**	0.28	0.30	0.31
	0.74	**0.68**	**0.73**	**0.62**	**0.69**	**0.53**
aw	**0.47**	**0.49**	0.43	**0.46**	**0.49**	**0.59**
	0.74	**0.73**	**0.76**	**0.74**	**0.63**	**0.61**
oy	**0.80**	**0.80**	**0.73**	**0.77**	**0.81**	**0.79**
	0.20	0.22	0.20	0.23	0.23	0.27

Table B.77: Correlation between perceptual scores and weighted distance measures computed using **nine** frames of **MFCC**s in **open set (5+5)** experiment

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	0.19	0.21	-0.00	0.07	0.02	0.17
	0.15	0.12	0.17	0.13	0.25	0.06
ow	0.20	0.14	0.02	-0.03	0.23	0.06
	-0.07	0.08	0.03	0.17	-0.02	0.31
ay	0.21	0.18	0.15	0.03	-0.16	0.10
	0.73	**0.68**	**0.62**	**0.53**	**0.61**	0.28
aw	0.32	0.38	0.26	0.36	0.34	**0.46**
	0.58	**0.54**	**0.58**	0.44	**0.51**	0.35
oy	**0.76**	**0.79**	**0.58**	**0.67**	**0.78**	**0.72**
	0.16	0.25	0.20	0.22	0.19	0.25

Table B.78: Correlation between perceptual scores and weighted distance measures computed using **nine** frames of **MFCC**s in **open set (5+5)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	**0.58**	**0.56**	**0.45**	**0.47**	**0.54**	0.44
	0.75	**0.67**	**0.74**	**0.59**	**0.72**	**0.56**
ow	0.26	0.27	0.34	0.34	0.27	0.24
	0.41	0.38	0.38	0.32	0.43	0.39
ay	0.39	0.38	0.35	0.29	**0.45**	0.30
	0.78	**0.76**	**0.73**	**0.74**	**0.70**	**0.76**
aw	0.09	0.12	0.12	0.09	0.14	0.25
	0.83	**0.81**	**0.79**	**0.73**	**0.83**	**0.78**
oy	0.32	0.23	0.33	0.18	0.26	0.22
	0.26	0.31	0.28	0.36	0.27	0.35

Table B.79: Correlation between perceptual scores and weighted distance measures computed using **nine** frames of **MFCC**s in **closed set** experiment

Diphthong	*Euclidean*		*Absolute*		*Mahalanobis*	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	**0.59**	**0.63**	**0.46**	**0.60**	**0.57**	**0.48**
	0.77	**0.70**	**0.79**	**0.63**	**0.73**	**0.51**
ow	0.30	0.30	**0.47**	0.44	0.35	0.29
	0.37	0.32	0.35	0.24	0.39	0.27
ay	0.32	0.34	0.27	0.24	**0.45**	0.36
	0.70	**0.69**	**0.68**	**0.72**	**0.67**	**0.71**
aw	-0.04	0.04	0.06	0.07	0.10	0.35
	0.81	**0.76**	**0.71**	**0.62**	**0.84**	**0.73**
oy	0.24	0.13	0.30	0.10	0.22	0.14
	0.30	0.38	0.30	0.42	0.27	0.41

Table B.80: Correlation between perceptual scores and weighted distance measures computed using **nine** frames of **MFCC**s in **closed set** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.15	0.27	0.23	0.33	0.36	0.24	0.38
	0.76	**0.62**	**0.77**	**0.63**	**0.74**	**0.74**	**0.47**
ow	0.42	0.25	0.39	0.26	0.33	**0.47**	0.21
	0.45	**0.52**	0.40	**0.45**	0.33	0.38	0.32
ay	0.05	-0.06	0.08	-0.06	0.35	0.08	-0.11
	0.53	**0.66**	**0.51**	**0.63**	**0.62**	**0.53**	**0.61**
aw	0.35	0.29	0.28	0.31	0.18	0.39	0.39
	0.82	**0.82**	**0.80**	**0.82**	**0.80**	**0.80**	**0.72**
oy	0.39	0.17	0.25	0.20	0.33	0.39	0.26
	0.13	0.22	0.16	0.21	0.32	0.17	0.28

Table B.81: Correlation between perceptual scores and weighted distance measures computed using **nine** frames of **LSF**s in **open set (9+1)** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.04	0.24	0.20	0.31	0.27	0.10	0.35
	0.74	**0.60**	**0.77**	**0.55**	**0.73**	**0.77**	**0.47**
ow	**0.50**	0.13	0.41	0.17	0.34	**0.52**	0.10
	0.46	**0.56**	0.40	**0.51**	0.29	0.40	0.35
ay	-0.00	-0.14	0.08	-0.16	0.35	0.05	-0.21
	0.24	**0.50**	0.38	**0.49**	0.39	0.40	**0.51**
aw	0.44	**0.46**	0.29	**0.53**	0.17	0.35	**0.58**
	0.75	**0.74**	**0.68**	**0.73**	**0.78**	**0.75**	**0.73**
oy	0.47	0.37	0.22	0.40	0.38	0.37	0.40
	0.08	0.21	0.17	0.20	0.29	0.15	0.18

Table B.82: Correlation between perceptual scores and weighted distance measures computed using **nine** frames of **LSF**s in **open set (9+1)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.15	0.22	0.16	0.21	0.33	0.18	0.27
	0.10	0.23	0.19	0.31	0.35	0.26	0.36
ow	**0.73**	**0.61**	**0.75**	**0.63**	**0.71**	**0.73**	**0.46**
	0.41	0.25	0.38	0.25	0.31	**0.45**	0.21
ay	**0.46**	**0.52**	0.41	**0.46**	0.33	0.39	0.35
	0.06	-0.05	0.10	-0.04	0.36	0.11	-0.11
aw	**0.53**	**0.66**	**0.49**	**0.62**	**0.69**	**0.53**	**0.61**
	0.39	0.31	0.33	0.35	0.28	0.41	**0.50**
oy	**0.82**	**0.83**	**0.80**	**0.83**	**0.80**	**0.80**	**0.73**
	0.38	0.16	0.25	0.18	0.33	0.38	0.22

Table B.83: Correlation between perceptual scores and weighted distance measures computed using **nine** frames of **LSF**s in **open set (8+2)** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.10	0.21	0.16	0.19	0.32	0.16	0.18
	0.04	0.21	0.18	0.29	0.28	0.13	0.31
ow	**0.73**	**0.60**	**0.74**	**0.58**	**0.67**	**0.76**	**0.50**
	0.42	0.12	0.35	0.15	0.30	**0.47**	0.11
ay	**0.47**	**0.53**	0.40	**0.50**	0.27	**0.42**	0.39
	0.02	-0.11	0.12	-0.12	0.36	0.10	-0.20
aw	0.26	**0.48**	0.40	**0.50**	**0.61**	0.32	**0.50**
	0.50	**0.47**	0.34	**0.53**	0.41	0.33	**0.66**
oy	**0.73**	**0.75**	**0.67**	**0.73**	**0.78**	**0.73**	**0.73**
	0.44	0.40	0.22	0.44	0.37	0.35	0.41

Table B.84: Correlation between perceptual scores and weighted distance measures computed using **nine** frames of **LSF**s in **open set (8+2)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.21	0.27	0.23	0.31	0.36	0.23	0.39
	0.37	0.04	0.39	0.15	0.23	0.42	0.03
ow	0.03	-0.16	0.08	-0.12	0.35	0.12	-0.17
	0.32	0.38	0.30	0.29	0.05	0.30	**0.48**
ay	0.32	0.02	0.22	0.04	0.25	0.37	0.11
	0.74	**0.64**	**0.76**	**0.65**	**0.75**	**0.73**	**0.51**
aw	**0.47**	**0.51**	0.40	0.42	0.38	0.39	0.32
	0.58	**0.67**	**0.62**	**0.72**	**0.64**	**0.56**	**0.60**
oy	**0.79**	**0.82**	**0.78**	**0.81**	**0.80**	**0.77**	**0.72**
	0.11	0.24	0.19	0.24	0.31	0.20	0.32

Table B.85: Correlation between perceptual scores and weighted distance measures computed using **nine** frames of **LSF**s in **open set (5+5)** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.05	0.21	0.13	0.24	0.23	0.06	0.31
	0.42	-0.02	0.37	0.04	0.16	0.32	-0.03
ow	-0.01	-0.15	0.00	-0.15	0.34	0.11	-0.15
	0.40	0.31	0.26	0.26	-0.12	0.27	0.33
ay	**0.47**	-0.01	0.32	-0.02	0.22	0.36	-0.04
	0.74	**0.51**	**0.75**	**0.49**	**0.77**	**0.74**	**0.52**
aw	**0.48**	**0.50**	0.38	0.38	0.33	0.36	0.36
	0.49	**0.57**	**0.54**	**0.64**	0.35	**0.53**	**0.56**
oy	**0.78**	**0.74**	**0.72**	**0.75**	**0.76**	**0.78**	**0.73**
	0.05	0.22	0.22	0.22	0.26	0.15	0.25

Table B.86: Correlation between perceptual scores and weighted distance measures computed using **nine** frames of **LSF**s in **open set (5+5)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.17	0.27	0.25	0.34	0.41	0.28	0.39
	0.76	**0.56**	**0.77**	**0.58**	**0.73**	**0.74**	0.45
ow	0.43	0.24	0.41	0.23	0.34	**0.47**	0.21
	0.44	**0.51**	0.39	**0.47**	0.35	0.40	0.37
ay	0.09	-0.02	0.12	-0.01	0.37	0.11	-0.04
	0.60	**0.68**	**0.58**	**0.65**	**0.69**	**0.57**	**0.67**
aw	0.43	0.40	0.33	0.39	0.35	**0.47**	**0.45**
	0.83	**0.83**	**0.81**	**0.83**	**0.81**	**0.80**	**0.72**
oy	0.43	0.23	0.25	0.28	0.36	0.39	0.32
	0.11	0.24	0.15	0.27	0.33	0.17	0.33

Table B.87: Correlation between perceptual scores and weighted distance measures computed using **nine** frames of **LSF**s in **closed set** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.09	0.29	0.24	0.34	0.38	0.18	0.37
	0.76	**0.59**	**0.78**	**0.55**	**0.76**	**0.78**	0.45
ow	**0.57**	0.14	**0.46**	0.13	0.42	**0.53**	0.09
	0.49	**0.62**	0.44	**0.58**	0.36	**0.45**	**0.53**
ay	0.11	-0.08	0.17	-0.09	0.38	0.14	-0.13
	0.47	**0.58**	**0.49**	**0.56**	**0.62**	**0.50**	**0.61**
aw	**0.57**	**0.71**	0.41	**0.70**	**0.52**	**0.51**	**0.71**
	0.79	**0.78**	**0.73**	**0.77**	**0.79**	**0.78**	**0.74**
oy	**0.49**	0.42	0.29	0.47	0.39	0.41	0.47
	0.09	0.25	0.23	0.30	0.34	0.20	0.27

Table B.88: Correlation between perceptual scores and weighted distance measures computed using **nine** frames of **LSF**s in **closed set** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.26	0.14	0.38	0.23	0.43	0.36	**0.47**
	0.61	**0.50**	**0.70**	**0.50**	**0.63**	**0.54**	**0.50**
ow	**0.48**	0.31	0.42	0.35	0.32	0.31	0.19
	0.51	0.39	**0.51**	0.41	**0.46**	**0.53**	0.41
ay	-0.01	0.13	-0.05	0.10	0.14	0.02	-0.01
	0.54	0.22	**0.54**	0.25	**0.67**	**0.59**	**0.60**
aw	0.31	0.21	0.28	0.27	0.22	**0.53**	**0.55**
	0.73	**0.55**	**0.75**	**0.54**	**0.74**	**0.73**	**0.76**
oy	0.09	**0.49**	0.09	**0.54**	0.06	0.25	0.36
	0.56	**0.46**	**0.50**	**0.57**	**0.49**	**0.52**	**0.73**

Table B.89: Correlation between perceptual scores and weighted distance measures computed using **nine** frames of **MCA**s in **open set (9+1)** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	-0.03	0.02	0.25	0.10	**0.49**	**0.54**	**0.47**
	0.48	0.46	**0.66**	**0.47**	**0.59**	**0.51**	0.37
ow	-0.00	0.11	-0.03	0.19	0.23	0.15	-0.05
	0.56	0.27	**0.59**	0.12	**0.57**	**0.57**	0.39
ay	-0.09	0.27	-0.14	0.22	0.05	-0.02	-0.05
	0.43	0.15	**0.52**	0.17	**0.59**	**0.60**	**0.55**
aw	-0.04	0.25	0.12	0.28	0.10	**0.47**	**0.52**
	0.48	**0.60**	**0.46**	**0.54**	**0.55**	**0.70**	**0.70**
oy	-0.08	0.38	-0.11	**0.51**	-0.08	0.05	0.38
	0.41	-0.10	0.30	0.02	0.35	0.28	**0.57**

Table B.90: Correlation between perceptual scores and weighted distance measures computed using **nine** frames of **MCA**s in **open set (9+1)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	**0.55**	**0.46**	**0.48**	**0.58**	**0.47**	**0.53**	**0.72**
	0.35	0.13	0.38	0.22	**0.51**	0.34	**0.47**
ow	**0.65**	**0.50**	**0.69**	**0.49**	**0.61**	**0.54**	**0.47**
	0.46	0.32	0.39	0.33	0.28	0.29	0.16
ay	**0.52**	0.41	**0.51**	0.40	**0.47**	**0.52**	0.40
	-0.02	0.14	-0.06	0.11	0.14	0.01	-0.04
aw	**0.55**	0.21	**0.53**	0.24	**0.67**	**0.58**	**0.60**
	0.31	0.22	0.27	0.28	0.23	**0.53**	**0.54**
oy	**0.73**	**0.55**	**0.73**	**0.54**	**0.72**	**0.76**	**0.75**
	0.10	**0.48**	0.10	**0.53**	0.07	0.28	0.34

Table B.91: Correlation between perceptual scores and weighted distance measures computed using **nine** frames of **MCA**s in **open set (8+2)** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.35	-0.02	0.28	0.06	0.30	0.31	0.44
	0.31	0.05	0.45	0.13	**0.65**	**0.59**	**0.48**
ow	**0.47**	0.45	0.31	0.41	**0.52**	**0.50**	0.36
	-0.04	0.07	-0.11	0.04	0.15	0.14	-0.06
ay	**0.57**	0.36	**0.60**	0.23	**0.57**	**0.55**	0.38
	-0.06	0.29	-0.13	0.28	0.07	-0.01	-0.06
aw	0.32	0.14	**0.46**	0.16	**0.58**	**0.49**	**0.50**
	-0.12	0.25	0.07	0.28	0.07	0.40	**0.47**
oy	**0.48**	**0.58**	0.45	**0.53**	**0.53**	**0.71**	**0.68**
	-0.11	0.30	-0.17	**0.45**	-0.10	0.05	0.38

Table B.92: Correlation between perceptual scores and weighted distance measures computed using **nine** frames of **MCA**s in **open set (8+2)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.22	0.15	**0.49**	0.24	0.40	0.41	0.38
	0.45	0.34	0.40	0.37	0.43	0.43	0.33
ow	-0.02	0.03	-0.07	0.04	0.11	0.01	0.00
	0.37	0.18	0.43	0.24	0.37	**0.60**	**0.64**
ay	-0.02	**0.49**	-0.02	**0.48**	-0.03	0.22	0.37
	0.59	**0.51**	**0.66**	**0.52**	**0.60**	**0.56**	0.43
aw	0.39	**0.46**	**0.45**	**0.45**	**0.47**	**0.45**	0.35
	0.60	0.23	**0.60**	0.25	**0.69**	**0.61**	**0.67**
oy	**0.74**	**0.56**	**0.74**	**0.55**	**0.74**	**0.72**	**0.74**
	0.59	**0.46**	**0.52**	**0.56**	**0.56**	**0.54**	**0.66**

Table B.93: Correlation between perceptual scores and weighted distance measures computed using **nine** frames of **MCA**s in **open set (5+5)** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.14	0.29	0.39	0.26	**0.50**	**0.55**	**0.47**
	0.53	0.18	0.35	0.24	**0.46**	0.29	0.28
ow	-0.13	-0.06	-0.15	0.03	0.01	-0.01	-0.02
	0.31	0.21	**0.50**	0.29	0.40	**0.56**	**0.64**
ay	-0.22	0.21	-0.29	0.30	-0.26	0.04	0.33
	0.36	0.36	**0.48**	0.40	**0.47**	0.41	0.24
aw	**0.50**	0.25	**0.53**	-0.12	**0.54**	**0.50**	0.38
	0.46	0.09	**0.49**	0.07	**0.56**	**0.59**	**0.56**
oy	**0.47**	0.37	0.38	0.35	**0.48**	**0.69**	**0.65**
	0.45	-0.29	0.19	-0.20	0.36	0.22	**0.53**

Table B.94: Correlation between perceptual scores and weighted distance measures computed using **nine** frames of **MCA**s in **open set (5+5)** experiment (with intercept)

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.21	0.17	0.32	0.25	**0.49**	0.39	0.42
	0.63	**0.52**	**0.71**	**0.51**	**0.64**	**0.54**	**0.50**
ow	**0.50**	0.34	**0.48**	0.33	0.40	0.38	0.28
	0.50	0.41	**0.50**	0.44	**0.48**	**0.51**	**0.55**
ay	-0.01	0.11	-0.05	0.10	0.14	0.02	-0.00
	0.55	0.24	**0.56**	0.27	**0.69**	**0.61**	**0.60**
aw	0.32	0.23	0.29	0.30	0.26	**0.53**	**0.57**
	0.75	**0.56**	**0.76**	**0.55**	**0.77**	**0.74**	**0.76**
oy	0.45	**0.53**	0.37	**0.58**	0.18	0.42	0.41
	0.40	**0.46**	0.38	**0.57**	**0.47**	**0.48**	**0.72**

Table B.95: Correlation between perceptual scores and weighted distance measures computed using **nine** frames of **MCA**s in **closed set** experiment

Diphthong	*Euclidean*		*Absolute*		*K-L*	*Mahalanobis*	
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.35	0.15	**0.50**	0.19	**0.68**	**0.62**	**0.48**
	0.51	**0.51**	**0.68**	**0.50**	**0.62**	**0.51**	0.38
ow	**0.53**	0.42	**0.47**	**0.46**	0.45	0.36	0.35
	0.62	**0.63**	**0.63**	**0.49**	**0.61**	**0.61**	**0.61**
ay	-0.09	0.33	-0.14	0.28	0.05	0.00	-0.03
	0.50	0.22	**0.57**	0.21	**0.67**	**0.64**	**0.57**
aw	0.18	0.27	0.34	0.32	0.27	**0.57**	**0.62**
	0.62	**0.61**	**0.58**	**0.56**	**0.64**	**0.74**	**0.74**
oy	0.13	**0.58**	-0.03	**0.65**	0.01	0.18	0.44
	0.47	0.11	0.43	0.26	0.41	0.36	**0.65**

Table B.96: Correlation between perceptual scores and weighted distance measures computed using **nine** frames of **MCA**s in **closed set** experiment (with intercept)

Bibliography

Adell, J. (2003), Clustering linear dynamic models for the use of context-dependent models on speech recognition, Master's thesis, University of Edinburgh/UPC Barcelona.

Allen, J., Hunnicutt, S., Carlson, R. & Gramström, B. (1979), 'MITalk-79: the MIT text-to-speech system', *J. Acoust. Soc. Am.* **65**(1), S130.

Allen, J., Hunnicutt, S. & Klatt, D. H. (1987), *From Text to Speech: the MITalk System*, Cambridge University Press, UK.

Alvarez, Y. V. & Huckvale, M. (2002), The reliability of the ITU-T P.85 standard for the evaluation of the text-to-speech systems, *in* 'Proc. ICSLP', Denver, USA.

Beutnagel, M., Conkie, A., Schroeter, J., Stylianou, Y. & Syrdal, A. (1999), The AT&T Next-Gen TTS system, *in* 'Proc. Joint Meeting of ASA, EAA, and DEGA', Berlin, Germany.

Bishop, C. M. (1995), *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, UK.

Black, A. (2002), Perfect synthesis for all of the people all of the time, *in* 'Proc. IEEE 2002 Workshop on Speech Synthesis', Santa Monica, USA.

Black, A. & Campbell, N. (1995), Optimising selection of units from speech databases for concatenative synthesis, *in* 'Proc. Eurospeech', pp. 581–584.

Black, A. & Taylor, P. (1994), CHATR: A generic speech synthesis system, *in* 'Proc. COLING', pp. 983–986.

Black, A. & Taylor, P. (1997*a*), Automatically clustering similar units for unit selection in speech synthesis, *in* 'Proc. Eurospeech', Vol. 2, Rhodes, Greece, pp. 601–604.

Black, A. & Taylor, P. (1997*b*), The Festival speech synthesis system: system documentation, Technical Report HCRC/TR-83, Human Communication Research Centre, Univ. of Edinburgh, Edinburgh, Scotland.

Blouin, C., Rosec, O., Bagshaw, P. & d'Alessandro, C. (2002), Concatenation cost calculation and optimisation for unit selection in TTS, *in* 'Proc. IEEE 2002 Workshop on Speech Synthesis', Santa Monica, USA.

Breen, A. P. & Jackson, P. (1998*a*), Non-uniform unit selection and the similarity metric within BT's Laureate TTS system, *in* 'Proc. 3rd ESCA workshop in Speech Synthesis', pp. 201–206.

Breen, A. P. & Jackson, P. (1998*b*), A phonologically motivated method of selecting non-uniform units, *in* 'Proc. ICSLP', Sydney, Australia.

Brieman, L., Friedman, J., Olshen, R. & Stone, C. (1992), *Classification and Regression Trees*, Wadsworth & Brooks, Pacific Grove, CA, USA.

Campbell, N. (1996), CHATR: A high-definition speech re-sequencing system, *in* 'Proc. 3rd ASA/ASJ Joint Meeting', Hawaii, pp. 1223–1228.

Chappell, D. T. & Hansen, J. H. (2002), 'A comparison of spectral smoothing methods for segment concatenation based speech synthesis', *Speech Communications* **36**, 343–374.

Charpentier, F. & Stella, M. (1986), Diphone synthesis using an overlap-add technique for speech waveform concatenation, *in* 'Proc. ICASSP', pp. 2015–2018.

Chen, J.-D. & Campbell, N. (1999), Objective distance measures for assessing concatenative speech synthesis, *in* 'Proc. Eurospeech', Budapest, Hungary.

Chu, M. & Peng, H. (2001), An objective measure for estimating MOS of synthesized speech, *in* 'Proc. Eurospeech', Aalborg, Denmark.

Conkie, A., Beutnagel, M., Syrdal, A. & Brown, P. (2000), Preselection of a candidate units in a unit selection-based text-to-speech synthesis system, *in* 'Proc. ICSLP', Beijing, China.

Conkie, A. & Isard, S. (1996), Optimal coupling of diphones, *in* J. van Santen, R. Sproat, J. Olive & J. Hirschberg, eds, 'Progress in speech synthesis', Springer Verlag, pp. 293–305.

Coorman, G., Fackrell, J., Rutten, P. & van Coile, B. (2000), Segment selection in the L & H RealSpeak laboratory TTS system, *in* 'Proc. ICSLP', Beijing, China.

Crowe, A. & Jack, M. A. (1987), 'Globally optimising formant tracker using generalised centroids', *Electronic Letters* **23**(19), 1019–1020.

Davis, S. B. & Mermelstein, P. (1980), 'Comparison of parameteric representations for monosyllabic word recognition in continuously spoken sentences', *IEEE Trans. Acoustics, Speech, Signal Processing* **28**(4), 357–366.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society Series B* **39**, 1–38.

Digilakis, V., Rohlicek, J. & Ostendorf, M. (1993), 'ML estimation of a stochastic linear system with the EM algorithm and its applications to speech recognition', *IEEE Trans. Speech and Audio Processing* **1**(4), 431–442.

Ding, W. & Campbell, N. (1997), Optimising unit selection with voice source and formants in the CHATR speech synthesis system, *in* 'Proc. Eurospeech', Rhodes, Greece.

Ding, W., Kasuya, H. & Adachi, S. (1995), 'Simultaneous estimation of vocal tract and voice source parameters based on ARX model', *IEICE Trans. Inf. & Syst.* **E78-D**.

Donovan, R. E. (2001), A new distance measure for costing spectral discontinuities in concatenative speech synthesisers, *in* 'The 4th ISCA Tutorial and Research Workshop on Speech Synthesis', Perthshire, Scotland, pp. 59–62.

Donovan, R. E. (2003), 'Topics in decision tree based speech synthesis', *Computer Speech and Language* **17**, 43–67.

Donovan, R. E. & Eide, E. M. (1998), The IBM trainable speech synthesis system, *in* 'Proc. ICSLP', Sydney, Australia.

Donovan, R. E., Ittycheriah, A., Franz, M., Ramabhadran, B., Eide, E., Viswanathan, M., Bakis, R., Hamza, W., Picheny, M., Gleason, P., Rutherfoord, T., Cox, P., Green, D., Janke, E., Revelin, S., Waast, C., Zeller, B., Guenther, C. & Kunzmann, J. (2001), Current status of the IBM trainable speech synthesis system, *in* 'The 4th ISCA Tutorial and Research Workshop on Speech Synthesis', Perthshire, Scotland.

Dutoit, T. (1997), *An Introduction to Text-to-Speech Synthesis*, Kluwer Academic Publishers, The Netherlands.

Dutoit, T. & Leich, H. (1993), 'MBR-PSOLA: Text-to-speech synthesis based on an MBE re-synthesis of the segments database', *Speech Communications* **13**(3-4), 435–440.

Erkelens, J. & Broersen, P. (1995), On the statistical properties of line spectrum pairs, *in* 'Proc. ICASSP', Detroit, USA, pp. 768–771.

Founda, M., Tambouratzis, G., Chalamandaris, A. & Carayannis, G. (2001), Reducing spectral mismatches in concatenative speech synthesis via systematic database enrichment, *in* 'Proc. Eurospeech', Aalborg, Denmark.

Frankel, J. (2003), Linear dynamic models for automatic speech recognition, PhD thesis, University of Edinburgh.

Frankel, J. & King, S. (2001), ASR - articulatory speech recognition, *in* 'Proc. Eurospeech', Aalborg, Denmark, pp. 599–602.

Furui, S. (1986), 'Speaker-independent isolated word recognition using dynamic features of speech spectrum', *IEEE Trans. Acoustics, Speech, Signal Processing* **34**, 52–59.

Ghahramani, Z. & Hinton, G. (1996), Parameter estimation for linear dynamical systems, *in* 'Tech. rep. CRG-TR-96-2', Dept. of Computer Science, Univ. of Toronto. Software at www.gatsby.ucl.ac.uk/~ zoubin/software.html.

Gray, A. H., J. & Markel, J. D. (1976), 'Distance measures for speech processing', *IEEE Trans. Acoustics, Speech, Signal Processing* **24**(5), 380–391.

Hamon, C., Moulines, E. & Charpentier, F. (1989), A diphone system based on time-domain prosodic modifications of speech, *in* 'Proc. ICASSP', Vol. S5.7, pp. 238–241.

Hermansky, H. (1990), 'Perceptual linear predictive (PLP) analysis of speech', *J. Acoust. Soc. Am.* **87**(4), 1738–1752.

Huang, X., Acero, A., Adcock, J., Hon, H.-W., Goldsmith, J., Liu, J. & Plumpe, M. (1996), Whistler: A trainable text-to-speech system, *in* 'Proc. ICSLP', Philadelphia, USA, pp. 2387–2390.

Huang, X., Acero, A. & Hon, H.-W. (2001), *Spoken Language Processing, A guide to Theory , Algorithms and Sysystem Development*, Prentice Hall, New Jersey, USA.

Hunt, A. & Black, A. (1996), Unit selection in a concatenative speech synthesis system using a large speech database, *in* 'Proc. ICASSP', pp. 373–376.

Hunt, M., Zwierynski, D. & Carr, R. (1989), Issues in high quality LPC analysis and synthesis, *in* 'Proc. Eurospeech', pp. 348–351.

Itakura, F. (1975), 'Line spectrum representation of linear predictor coefficients of speech signals', *J. Acoust. Soc. Am.* **57**, S35(A).

Iwahashi, N., Kaiki, N. & Sagisaka, Y. (1992), Concatenative speech synthesis by minimum distortion criteria, *in* 'Proc. ICASSP', Vol. 2, pp. 65–68.

Kalman, R. E. (1960), 'A new approach to linear filtering and prediction problems', *Trans. Am.Soc.Mech.Eng., Series D, Journal of Basic Engineering* **82**, 35–45.

Kawai, H. & Tsuzaki, M. (2002), Acoustic measures vs phonetic features as predictors of audible discontinuity in concatenative synthesis, *in* 'Proc. ICSLP', Denver, USA.

Klabbers, E. & Veldhuis, R. (1998), On the reduction of concatenation artefacts in diphone synthesis, *in* 'Proc. ICSLP', Vol. 6, Sydney, Australia, pp. 1983–1986.

Klabbers, E. & Veldhuis, R. (2001), 'Reducing audible spectral discontinuities', *IEEE Trans. Speech and Audio Processing* **9**(1), 39–51.

Klabbers, E., Veldhuis, R. & Koppen, K. (2000), A solution to the reduction of concatenation artefacts in speech synthesis, *in* 'Proc. ICSLP', Beijing, China.

Klatt, D. H. (1987), 'Review of text-to-speech conversion for English', *J. Acoust. Soc. Am.* **82**(3), 737–793.

Kullback, S. & Leibler, R. (1951), 'On information and sufficiency', *Annals of Mathematical Statistics* **22**, 79–86.

Lee, M. (2001), Perceptual cost functions for unit searching in large corpus-based concatenative text-to-speech, *in* 'Proc. Eurospeech', Aalborg, Denmark.

Lee, M., Lopresti, D. P. & Olive, J. P. (2001), A text-to-speech platform for variable length optimal unit searching using perceptual cost functions, *in* 'The 4th ISCA Tutorial and Research Workshop on Speech Synthesis', Perthshire, Scotland, pp. 75–80.

Lenzo, K. A. & Black, A. (2000), Diphone collection and synthesis, *in* 'Proc. ICSLP', Beijing, China.

Makhoul, J. (1973), 'Spectral analysis of speech by linear prediction', *IEEE Trans. Audio and Electroacoustics* **21**(3), 140 –148.

Makhoul, J. (1975), 'Linear prediction: A tutorial review', *Proc. IEEE* **63**(4), 561 –580.

McGhee, W. J. (1985), *Introductory Statistics*, West Publishing Company, St. Paul, USA.

McKenna, J. (2001), Automatic glottal closed-phase location and analysis by Kalman filtering, *in* 'The 4th ISCA Tutorial and Research Workshop on Speech Synthesis', Perthshire, Scotland, pp. 91–96.

McKenna, J. & Isard, S. (1999), Tailoring Kalman filtering towards speaker characterisation, *in* 'Proc. Eurospeech', Budapest, Hungary, pp. 2793–2796.

Moulines, E. & Charpentier, F. (1990), 'Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones', *Speech Communication* **9**(5-6), 453–467.

Paliwal, K. K. & Kleijn, W. B. (1995), Quantization of LPC parameters, *in* W. B. Kleijn & K. K. Paliwal, eds, 'Speech Coding and Synthesis', Elsevier, Amsterdam, The Netherlands, pp. 433–466.

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

Peng, H., Zhao, Y. & Chu, M. (2002), Perpetually optimizing the cost function for unit selection in a TTS system with one single run of MOS evaluation, *in* 'Proc. ICSLP', Denver, USA.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992), *Numerical Recipes in C - The Art of Scientific Computing*, Cambridge University Press, Cambridge, UK.

Rabiner, L. & Juang, B. (1993), *Fundamentals of Speech Recognition*, Prentice Hall, New Jersey, USA.

Rauch, H. E., Tung, F. & Striebel, C. T. (1965), 'Maximum likelihood estimates of linear dynamic systems', *AIAA Journal* **3**(8), 1445–1450.

Richmond, K. (2002), Estimating the articulatory parameters from the acoustic speech signal, PhD thesis, University of Edinburgh.

Rowies, S. & Ghahramani, Z. (1999), 'A unifying review of linear gaussian models', *Journal of Neural Computation* **11**(2).

Rutten, P., Aylett, M. P., Fackrell, J. & Taylor, P. (2002), A statistically motivated database pruning technique for unit selection synthesis, *in* 'Proc. ICSLP', Denver, USA.

Shumway, R. H. & Stoffer, D. S. (1982), 'An approach to time series smoothing and forecasting using the EM algorithm', *Journal of Time Series Analysis* **3**(4), 253–264.

Smith, G., de Frietas, J., Robinson, T. & Niranjan, M. (1999), 'Speech modelling using subspace and EM techniques', *Advances in Neural Information Processing Systems* **12**, 796–802.

Soong, F. K. & Juang, B. H. (1984), Line spectrum pairs (LSP) and speech data compression, *in* 'Proc. ICASSP', pp. 1.10.1–1.10.4.

Spiegel, M. R. (1975), *Theory and Problems of Probability and Statistics*, Schaum's Outline Series in Mathematics, McGraw-Hill Book Company.

Stylianou, Y. (2001), 'Applying the harmonic plus noise model in concatenative speech synthesis', *IEEE Trans. Speech and Audio Processing* **9**(1), 21–29.

Stylianou, Y., Dutoit, T. & Schroeter, J. (1997), Diphone concatenation using a harmonic plus noise model of speech, *in* 'Eurospeech', pp. 613–616.

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

Sugamura, N. & Itakura, F. (1986), 'Speech analysis and synthesis methods developed at ECL in NTT (from LPC to LSP)', *Speech Communications* **5**, 199–215.

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

Taylor, P. (2000), 'Concept-to-speech by phonological structure matching', *Philosophical Transactions of the Royal Society, Series A* .

Taylor, P. & Black, A. (1999), Speech synthesis by phonological structure matching, *in* 'Proc. Eurospeech', Budapest, Hungary.

Taylor, P., Black, A. & Caley, R. (1998), The architecture of the Festival speech synthesis system, *in* 'Proc. 3rd ESCA workshop in Speech Synthesis', pp. 147–151.

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

Umezaki, T. & Itakura, F. (1986), Analysis of time fluctuating characteristics of linear predictive coefficients, *in* 'Proc. ICASSP', pp. 1257–1260.

van Santen, J., Pols, C. W., Abe, M., Kahn, D., Keller, E. & Vonwiller, J. (1998), Report on the third ESCA TTS workshop evaluation procedure, *in* 'Proc. 3rd ESCA workshop in Speech Synthesis'.

van Santen, J., Wouters, J. & Kain, A. (2002), Modification of speech: A tribute to Mike Macon, *in* 'Proc. IEEE 2002 Workshop on Speech Synthesis', Santa Monica, USA.

Veldhuis, R. & Klabbers, E. (2003), 'On the computation of Kullback-Leibler measure for spectral distances', *IEEE Trans. Speech and Audio Processing* **11**(1), 100–103.

Vepa, J. & King, S. (2003), Kalman-filter based join cost for unit-selection speech synthesis, *in* 'Eurospeech', Geneva, Switzerland.

Vepa, J. & King, S. (2004), Subjective evaluation of join cost & smoothing methods, *in* 'Proc. 5th ISCA Speech Synthesis Workshop', Pittsburgh, USA.

Vepa, J. & King, S. (forthcoming), Join cost for unit selection speech synthesis, *in* A. Alwan & S. Narayanan, eds, 'Speech Synthesis', Prentice Hall.

Vepa, J., King, S. & Taylor, P. (2002*a*), New objective distance measures for spectral discontinuities in concatenative speech synthesis, *in* 'Proc. IEEE 2002 Workshop on Speech Synthesis', Santa Monica, USA.

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

Viterbi, A. J. (1967), 'Error bounds for convolutional codes and an asymptotically optimum decoding algorithm', *IEEE Trans. on Information Theory* **13**(2), 260–269.

Wouters, J. & Macon, M. (1998), Perceptual evaluation of distance measures for concatenative speech synthesis, *in* 'Proc. ICSLP', Vol. 6, Sydney, Australia, pp. 2747–2750.

Wrench, A. (2001), A new resource for production modelling in speech technology", *in* 'Proc. Workshop on Innovations in Speech Processing', Stratford-upon-Avon, UK.

Wrench, A. A. (1995), Analysis of fricatives using multiple centres of gravity, *in* 'Proc. International Congress of Phonetic Sciences', Vol. 4, pp. 460–463.

Wrench, A. A., Watson, J. M. M., Soutar, D. S., Robertson, A. G. & Laver, J. (1994), Fast formant estimation of children's speech, *in* 'Proc. ICSLP', Vol. 3, pp. 1651–1654.

Young, S., Odell, J., Ollason, D., Valtchev, V. & Woodland, P. (1997), *The HTK Book*, Version 2.1, Cambridge University, Entropic Cambridge Research Laboratory, UK.