University of Crete
Department of Computer Science

# Detection of Discontinuities in Concatenative Speech Synthesis

(MSc. Thesis)

**Ioannis Pantazis**

Heraklion

Fall 2006

Department of Computer Science

University of Crete

# Detection of Discontinuities in Concatenative Speech Synthesis

Submitted to the

Department of Computer Science

in partial fulfillment of the requirements for the degree of

Master of Science

Fall, 2006

Author:

Ioannis Pantazis

Department of Computer Science

Board
of enquiry:

Supervisor

Ioannis Stylianou

Associate Professor

Member

Georgios Tziritas

Professor

Member

Panagiotis Tsakalides

Associate Professor

Accepted by:

Chairman of the
Graduate Studies Committee

Panagiotis Trahanias

Professor

Heraklion, Fall 2006

# Abstract

Last decade, unit selection synthesis became a hot topic in speech synthesis research. Unit selection gives the greatest naturalness due to the fact that it does not apply a large amount of digital signal processing to the recorded speech, which often makes recorded speech sound less natural. In order to find the best units in the database, unit selection is based on two cost functions, *target cost* and *concatenation cost*.

concatenation cost refers to how well adjacent units can be joined. The problem of finding a concatenation cost function is broken into two subproblems; into finding the proper parameterizations of the signal and into finding the right distance measure. Recent studies attempted to specify which concatenation distance measures are able to predict audible discontinuities and thus, highly correlates with human perception of discontinuity at concatenation point. However, none of the concatenation costs used so far, can measure the similarity (or, (dis-)continuity) of two consecutive units efficiently.

Many features such as line spectral frequencies (LSF) and Mel frequency cepstral coefficients (MFCC) have been used for the detection of discontinuities. In this study, three new sets of features for detecting discontinuities are introduced. The first set of features are obtained by modeling the speech signal as a sum of harmonics with time varying complex amplitude, which yield a nonlinear speech model. The second set of features is based on a nonlinear speech analysis technique which tries to decompose speech signals into AM and FM components. The third feature set exploits the nonlinear nature of the ear. Using Lyon's auditory model, the behavior of the cochlea is measured by evaluating neural firing rates.

To measure the difference between two vectors of such parameters, we need a distance measure. Examples of such measures are absolute distance ($l_1$ norm) and Euclidean distance ($l_2$ norm). However, these measures are naive and provide rather poor results. We further suggest using Fisher's linear discriminant as well as a quadratic discriminant as discrimination functions. Linear

regression, which employs a least-squares method, was also tested as a discrimination function.

The evaluation of the objective distance measures (or concatenation costs) as well as the training of the discriminant functions was performed on two databases. To build a database, a psychoacoustic listening experiment is performed and listener's opinions are obtained. The first database was created by Klabbers and Veldhuis in Holland while, the second database was created by Stylianou and Syrdal at AT&T Labs. Therefore, we are able to compare same approaches on different databases and obtain more robust results.

Results obtained from the two different psychoacoustic listening tests showed that nonlinear harmonic model using Fisher's linear discriminant or linear regression performed very well in both tests. It was significantly better than MFCC separated with Euclidean distance which a common concatenation cost in modern TTS systems. Another good concatenation cost, but less good than nonlinear harmonic model, is AM-FM decomposition again with Fisher's linear discriminant or linear regression. These results indicate that **a concatenation cost which is based on nonlinear features separated by a statistical discriminant function** is a good choice.

# Acknowledgements

This thesis is the result of a two-year working experience as graduate student at the research group for Media Informatics Lab at Computer Science Department of the University of Crete. This work was supported by a graduate fellowship from the University of Crete, which I also truly acknowledge for providing the necessary technical equipment.

First of all I would like to thank Yannis Stylianou, my supervisor, for his guidance and encouragement during this work! These years of collaboration have been very instructive, giving rise to several interesting tasks more or less related to the research. I would like to thank also Esther Klabbers from OGI for providing us with her valuable experiment data.

The pleasant and creative research atmosphere at the group is of course also an outcome of all the other members of the group. Of these I want to give special thanks to Y. Sfakianakis, A. Holzapfel, M. Vasilakis and G. Tzagarakis (he also contributed in formatting part of the LaTeX document) for sharing with me some of their profound knowledge on digital signal processing, to N. Melikides, M. Fundulakis and L. Archoleon, my friends, for the discussions on just everything that is interesting in this world. I would like also to thank Euaggelia for supporting and tolerating me all this time.

Finally, I feel obliged to thank my family, for providing me with the ideal environment to grow up in, and exhorting me in changing for the better. Their constant support and encouragement have really brought me here.

I have had a lot of fun. Thank you all (I hope that I've not forgotten someone).

# Contents

# List of Tables

# List of Figures

# List of Abbreviations

| | |
|---|---|
| AGC | Automatic gain control |
| $a_k$ | Amplitude of nonlinear harmonic model |
| AM | Amplitude modulated |
| $b_k$ | Slope of nonlinear harmonic model |
| DESA | Discrete energy separation algorithm |
| FLD | Fisher's linear discriminant |
| FM | Frequency modulated |
| HWR | Half wave rectification |
| LPC | Linear prediction coefficients |
| LR | Linear regression |
| LSF | Line spectral frequencies |
| MFCC | Mel-scaled frequency cepstral coefficients |
| ROC | Receiver operating characteristic |
| TTS | Text-to-Speech |
| QD | Quadratic discriminant |

# Chapter 1

# Introduction

Speech is a distinctive feature of human beings and it is used primarily for communication. Speech not only is the oldest means of communication but also the most widely used. Artificial speech, i.e. speech that is generated in a automated way, has been a dream of the humankind for centuries. Speech which is so easy and natural to be produced by humans is extremely difficult to be synthesized by machines of any kind making speech a rather complex signal. The realization of speech machines or systems has been really practical within the last forty years [1].

## 1.1 First Steps in Speech Synthesis

Long before modern electronic signal processing was inverted, speech researchers tried to build machines to create human speech [2]. In 1779, Professor Kratzenstein built models of the human vocal tract and explained the difference between five long vowels (/$a$/, /$e$/, /$i$/, /$o$/, /$u$/). He made apparatus to produce them artificially. Wolfgang von Kempelen, few years later in 1791, described a machine which was able to produce vowels as well as consonants. He added models for tongue and lips and produced more complicated sounds. In 1837, Wheatstone constructed a mechanical "speaking machine" based on von Kempelen's design.

In the 1930s researchers at Bell Labs developed the VOCODER, a keyboard-operated elec-

tronic speech analyzer and synthesizer that was said to be "clearly" intelligible. Homer Dudley refined this device into the VODER, which he exhibited at the 1939 New York World's affair.

After demonstration of VODER the scientific world became more and more interested in speech synthesis, due to the fact that it was finally shown that intelligible speech can be produced artificially. Actually, the basic structure and idea of VODER is very similar to source/filter models of speech. People realized that the speech signal could be decomposed as a source/filter model, with the glottis acting as a sound source and the vocal track being a filter. This model was used to build analog electronic devices that could mimic human speech [1]. Early electronic speech synthesizers sounded very robotic and were often barely intelligible. However, the quality of synthesized speech has steadily improved, and output from contemporary speech synthesis systems is sometimes indistinguishable from natural speech [3], [4].

Using the source/filter model first formant synthesizer was introduced by Laurence in 1953. It consisted of three electronic formant resonators connected in parallel. Fant, at about the same time, constructed a formant synthesizer in cascade and ten years later created a formant synthesizer with separate parts to model the transfer function of the vocal track and various excitations [1], [5]. First articulatory synthesizer was introduced by Rosen in 1958 and it was the first fully computer-based speech synthesis system ever created.

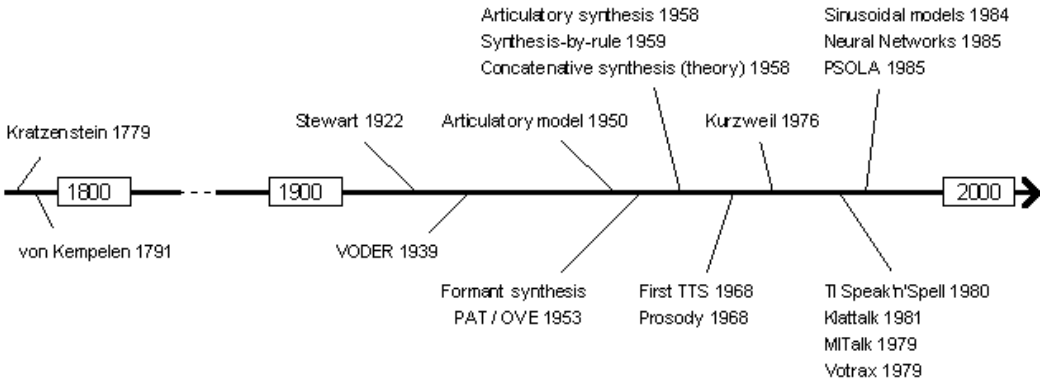



Figure 1.1: Evolution of Speech Synthesis Systems through the years. It is obvious that the advent of computers made speech synthesis to blossom.

However, much of the work in synthesis in 40s and 50s was primarily concerned with con-

structing replicas of the signal itself rather than generating the phones from an abstract form like text. In Figure 1.1, milestones of speech synthesizers through the years is showed.

## 1.2 Text-to-Speech Synthesis

In 1968, the first text-to-speech (TTS) system was developed by Umeda and his companions in Japan. It was based on an articulatory model for speech waveform production and a syntactic analysis module with sophisticated heuristics for text analysis. In general, TTS systems consist of two large modules (Figure 1.2). First module is *natural language processing* where morpho-syntactic analysis, phonetization and prosody generation take place and second module is *digital signal processing* where acoustic realization of each phoneme is yield [6], [7], [3].

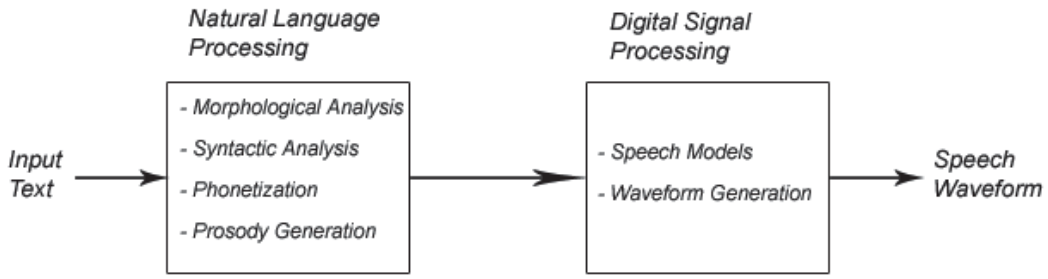



Figure 1.2: Modules of a Text-to-Speech Synthesis System. In order to produce high quality synthesized speech, experts from different fields should collaborate.

Our interest is concentrated on the second module where two different approaches for generating speech waveforms are used.

Firstly, the parametric approach where speech is modeled with a source/filter technique. Source/filter techniques such as articulatory or formant techniques and linear predictive coding produced speech which was somewhat artificial and robotic, yet, very intelligible and they became popular commercial products. In 1979, Klatt and colleagues demonstrated the MITalk synthesizer, a text-to-speech system developed at MIT. Two years later Klatt created a more

sophisticated TTS system called Klattalk. These systems in a sense defined the perception of automatic speech synthesis. Speak'n'Spell toy of Texas Instruments in the late 70s is an example –possibly the first– of a commercial TTS synthesis product [1].

The second approach, which displaced the parametric one, is based on the concatenation of speech segments (or units[1]) [8] [4]. With this method, speech is generated from recorded speech rather than being generated from models. As a consequence of this, output speech is sounding more natural.

## 1.3 Concatenative Speech Synthesis

The rise of concatenative synthesis began in the 70s, and has become practical in late 80s as large scale electronic storage has become cheap and robust. In concatenative synthesis, it is necessary to create a database with the desired speech elements. Depending on the contexts of the database, there are two main subtypes of concatenative synthesis, diphone synthesis and unit selection synthesis.

### 1.3.1 Diphone Synthesis

Diphone synthesis uses a minimal speech database containing all the diphones (phoneme[2] to phoneme transition) occurring in a given language [9], [10]. At runtime, target prosody of a sentence is superimposed on these minimal units by means of digital signal processing techniques such as LPC, PSOLA and MBROLA [3]. Diphone synthesis has the drawback that it suffers from sonic glitches at concatenation points and its advantage is the low storage size which nowadays is not a problem.

[1] In this thesis, words *segment* and *unit* are used interchangeably

[2] phoneme = the smallest unit of speech that differentiates one word from another

### 1.3.2 Unit Selection Synthesis

Unit selection synthesis was first demonstrated in middle 90s and quickly became a hot topic in speech synthesis research [11], [8]. It uses large speech databases with more than an hour of recorded speech. The primary motivation for the use of large databases is that with a large number of units with varied prosodic and spectral characteristics it should be possible to synthesize more natural-sounding speech than can be produced with a small set of controlled units. During database creation, each recorded utterance is segmented into single phones, diphones, triphones or even bigger units (such as words).

Typically, the division into segments is done using a specially modified speech recognizer set to a "forced alignment" mode with some hand correction afterward, using visual representations such as the waveform and spectrogram [10]. An index of the units in the speech database is then created based on the segmentation and acoustic parameters like the fundamental frequency (pitch), duration, position in the syllable, and neighboring phones. At runtime, the desired target utterance is created by determining the best chain of candidate units from the database (unit selection). The subject of this thesis, as it is explained below, is to find the best chain of units using objective evaluation criteria.

Unit selection gives the greatest naturalness due to the fact that it does not apply a large amount of digital signal processing to the recorded speech, which often makes recorded speech sound less natural, although some systems may use a small amount of signal processing at the point of concatenation to smooth the waveform [12], [13], [14]. In fact, output from the best unit-selection systems is often indistinguishable from real human voices. However, despite examples of it working excellently, generalized unit selection is known to producing bad quality synthesis from time to time. Our goal is to predict and eliminate such undesired cases.

More analytically, unit selection synthesis systems consists of five modules:

— *a text processing module,*

— *an acoustic- and prosodic-feature prediction module,*

— *a prerecorded speech database (corpus)*

— *a segment (unit) selection module and*

— *a waveform concatenation module.*

Figure 1.3 shows the flow diagram of a unit-selection synthesizer.

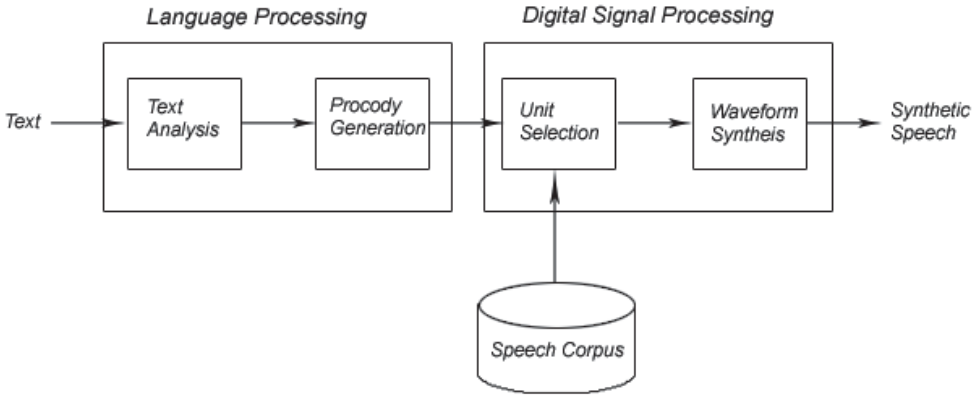



Figure 1.3: Modules of a Unit Selection Synthesis System.

## 1.4 Unit Selection Module

Prosody generation module returns the target specifications (or, simply target) for an utterance. The target defines the string of phonemes required to synthesize the text, and is annotated with prosodic features (i.e. duration, power and pitch) which specify the desired speech output in more detail [15]. Unit selection module has to search for the best units in the database (or corpus) given the target specifications.

Usually, unit selection is based on two cost functions shown in Figure 1.4. The target cost, $C_t(t_i, u_i)$, is an estimate of the difference between a database unit, $u_i$, and the target, $t_i$, which is supposed to represent. To put it in another way, target cost expresses the closeness between the context of the target and the candidate unit, therefore, it is calculated as a weighted sum of

the differences between prosodic and phonetic parameters.

The concatenation[3] cost, $C_c(u_{i-1}, u_i)$, is an estimate of the quality of a join between consecutive units. In other words, concatenation cost refers to how well adjacent units can be joined. It is calculated as a weighted sum of the differences between fundamental frequency, spectral mismatches, energy, etc.

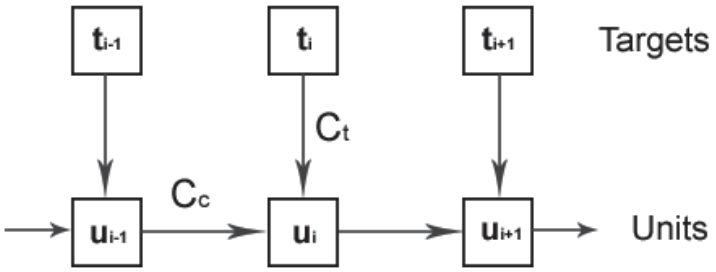



Figure 1.4: Target cost and concatenation cost. Target cost measures the closeness between target, $t_i$, and unit, $u_i$. Concatenation cost measures the continuity between previous unit, $u_{i-1}$ , and current unit, $u_i$.

Indeed, given the target specification, the sequence $\mathbf{t^n} = (t_1, t_2, ..., t_n)$, we have to select the set of units, $\mathbf{u^n} = (u_1, u_2, ..., u_n)$, which are closest to the target [15]. For each target $t_i$ there are more than one unit in the database. The different instances of each speech unit have various prosodic and spectral characteristics. For each unit, target cost and concatenation cost are computed and a network is constructed (Figure 1.5). Optimum unit selection is achieved by a Viterbi search for the lowest total cost path through the lattice of candidate units.

Among the two costs, the concatenation cost is more important for the selection of two successive acoustic units. In this thesis, we **studied and experimented on the calculation of the concatenation cost**.

Hunt and Black in their pioneering work [8] stated "*An objective distance measure should*

[3]Concatenation cost is also known as join cost

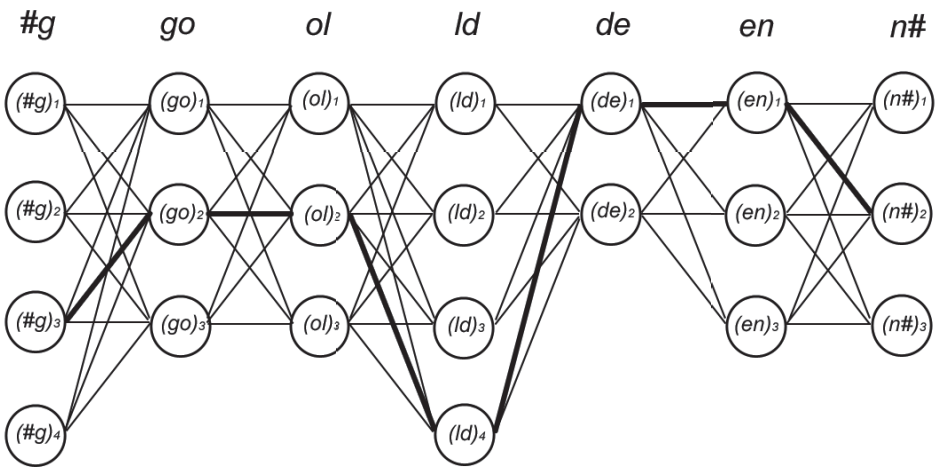



Figure 1.5: Candidate units constitute a k-partite graph. Each vertex has a weight which is the sum of target cost and concatenation cost. The shortest path is specified through Viterbi search algorithm. Here, best sequence of word *golden*, which is segmented into diphones, is found.

*reflect as much as possible the perceptual similarity of the utterances.*" Thus, the ideal concatenation cost is one that, although based solely on measurable properties of the candidate units —such as spectral parameters, amplitude, and F0— correlates highly with human listeners perceptions of discontinuity at concatenation points. In other words, the join cost should predict the degree of perceived discontinuity. Recent studies attempted to specify which concatenation distance measures are able to predict audible discontinuities and thus, highly correlates with human perception of discontinuity at concatenation point [16], [17], [18], [19], [20], [21], [22]. Having a "good" distance measure, high concatenation cost will be assigned to speech units that are identified to produce audible discontinuities. However, none of the concatenation costs used so far, can measure the similarity (or, (dis-)continuity) of two consecutive units efficiently.

Most of the previous studies computed concatenation cost by using a distance measure on some parameterizations of the speech signal (Figure 1.6). We used the same framework for two reasons. Firstly, using this approach the problem of finding a concatenation cost function is broken into two problems; into finding the proper parameterizations and into finding the right distance measure, and secondly, parameterizations of signals as well as distance measures have

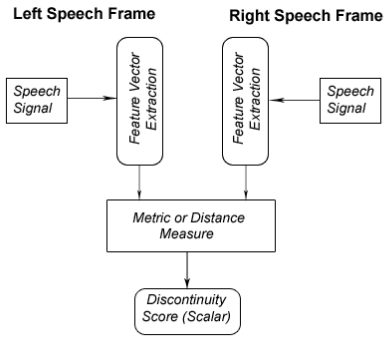



Figure 1.6: Flow diagram for measuring the discontinuity of two successive speech units.

meaningful physical interpretations.

Indeed, the concatenation cost contains a module that extracts the spectral properties of the speech from both candidate units. Many features such as line spectral frequencies (LSF) and Mel frequency cepstral coefficients (MFCC) have been used for the detection of discontinuities. In this study, three new sets of features for detecting discontinuities are introduced. Our goal is to increase the detection rate of perceived discontinuities, so, we suggest using features obtained from nonlinear approaches.

The first set of features are obtained by modeling the speech signal as a sum of harmonics with time varying complex amplitude [23], which yield a nonlinear speech model. The second set of features is based on a nonlinear speech analysis technique which tries to decompose speech signals into AM and FM components [24]. Speech signals pass through a filterbank which covers the most important frequencies of the speech spectrum, and then an algorithm referred to as DESA is applied for the separation of the AM and FM component. The third feature set exploits the nonlinear nature of the ear. Using Lyon's auditory model, the behavior of the cochlea is measured by evaluating neural firing rates. This model gives a two-dimensional representation of the signal which is similar to spectrogram, but it takes advantage of the nonlinear nature of the cochlea.

To measure the difference between two vectors of such parameters, we need a distance measure.

This may be a metric, provided that it has the required properties, but this is not necessary. Examples of such measures are absolute distance ($l_1$ norm) and Euclidean distance ($l_2$ norm). However, these measures are naive and provide rather poor results. We further suggest using Fisher's linear discriminant as well as a quadratic discriminant [25] as discrimination functions. Linear regression, which employs a least-squares method, was also tested as a discrimination function. Since these methods are statistical, the weights are learned from the data obtained from perceptual tests.

The evaluation of the objective distance measures as well as the training of the discriminant functions was performed on two databases. To build a database, a psychoacoustic listening experiment is performed and listener's opinions are obtained. The first database was created by Klabbers and Veldhuis [17] in Holland while, the second database was created by Stylianou and Syrdal [19] at AT&T Labs. Therefore, we are able to compare same approaches on different databases and obtain more robust results.

The results of this thesis were firstly presented in a summer school *"Nonlinear Speech Modeling and Applications"* in Italy [26]. Nonlinear harmonic model, AM-FM decomposition and Fisher's linear discriminant were introduced and evaluation was performed on AT&T's database. The results obtained from Klabbers's database were presented in *Interspeech 2005* on the same feature sets and discriminant functions [27]. Finally, a review on the problem of finding an objective distance measure for detecting audible discontinuities [28] was done for the same summer school *"Nonlinear Speech Modeling and Applications"* one year later in Heraklion.

## 1.5 Thesis Organization

In Chapter 2, literature is briefly reviewed and discussed, while in Chapter 3, we describe five different feature sets that were used in this study. Two of them have been already used for the detection of discontinuities while the other three are novel feature sets. Chapter 4 presents metrics

and statistical approaches for discriminating the signals with audible discontinuities from the non-audible ones. In Chapter 5, the databases where the evaluations of features and discrimination functions were performed, while in Chapter 6, results are shown. Finally, Chapter 7 concludes the methods and major results presented in this thesis.

# Chapter 2

# Literature Review

Clearly, unit selection has become a hot topic since Hunt and Black [8] introduced it. As a result of this, new challenges such as robust prosody generation, efficient construction of the speech database and detection of discontinuities showed up. For the problem we are interested, many researchers tried to solve it (i.e. to find an objective distance measure that is able to detect audible discontinuities). In order to obtain a global view, in this chapter, studies on the detection of discontinuities are briefly reviewed.

Wouters and Macon [16] in their study reported that a Euclidean distance on Mel-scale LPC-based cepstral parameters are o good predictor of perceived discontinuity. They evaluated several distance measures using perceptual data obtained from listening tests. In these tests, pairs of synthetic monosyllabic English words that were made from identical unit sequences, except for one half-phone, were presented to listeners. Then, their task was to rate the difference between the word pairs on a five-point scale. The substituted half-phones were limited to three specific vowels. They computed the correlation between objective distance measures and mean listener responses. Their results indicated that the parameterizations that use a nonlinear frequency scale, such as Mel or Bark scales, performed better than those that do not. They also found that weighting individual parameters of cepstra, LSF or delta coefficients could improve correlations.

As a continuation of their work, they presented an approach to reduce concatenation mismatches in concatenative speech synthesis [29]. They combined spectral information, represented by LSFs, from two sequences of speech units selected in parallel. The first sequence defined the initial spectral trajectories for a target utterance. Then, this sequence was modified by the second sequence which defined the desired transitions between concatenation units. Perceptual experiments showed that considerable amount of concatenation artifacts were removed.

Klabbers and Veldhuis [17] tested various spectral distance measures for joints in five Dutch vowels to find which measure best predicts the concatenation discontinuity. These various measures were correlated with the results of a listening experiment in which listeners have to make a choice between 0 or 1 based on whether the concatenation was perceived as smooth (0) or discontinuous (1). They found that Kullback-Leibler measure on LPC power-normalized spectra is the best predictor among their six spectral distance measures: Euclidean distances between first two formants, MFCC, the likelihood ratio, the mean squared log-spectral distance, loudness difference, and expectation differences.

They also studied the feasibility of extending a diphone inventory with context-sensitive diphones to reduce the occurrence of audible discontinuities [13]. In order to reduce the number of additional diphones, they used their best joint (concatenation) cost function to cluster the consonantal contexts that had the same spectral effects on neighboring vowels. To evaluate the improvements gained with this extended inventory, they conducted further perceptual experiments and observed that these additional diphones significantly reduced the number of audible discontinuities.

In order to identify the amount of perceptually noticeable spectral distortion between two speech segments, Hansen and Chappel introduced an auditory-based distance measure [30]. Their distortion measure uses Carney's computer model of the mammalian auditory system which produces realistic temporal response properties and average discharge rates of auditory nerves. They concluded that in combination with direct processing of time- and frequency-domain characteris-

tics, a small segment database using their distance measure achieves results close to concatenative systems with large databases. It is worth to note that their work was never used or commented by other researchers. We extend their work by using another auditory model —Lyon's auditory model— and make extensive evaluations.

A variety of acoustic transforms (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 compared by Chen and Campbell [18] for use in assessment and evaluation of synthetic speech. The speech material was synthesized using the CHATR speech synthesis system [31]. 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.

Stylianou and Syrdal conducted a psychoacoustic experiment on listeners detectability of signal discontinuities in concatenative speech synthesis. They used an experimental version of the AT&T next-generation system [19] to synthesize 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 distributions, 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 had the highest prediction rates. In contrast to [17], this study found that KL distance based on LPC spectra was the one of the worst performers.

Donovan [20] proposed a new join cost that can be described as a decision tree- based, context-dependent Mahalanobis distance on perceptual cepstral coefficients. He conducted listening tests to compare the performance of this new method with other joint 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 synthesized in a male voice using a modified form of the IBM trainable speech synthesis system [32]. The correlation results showed that this new measure outperforms other measures. Also, further listening tests have justified the use of this measure in the IBM synthesis system.

Vepa, King and Taylor [21] reported a perceptual experiment conducted to measure the correlation between subjective human perception and various objective spectral measures. They tested features such as LSF, MFCC and multiple centroid analysis (MCA) and metrics such as Euclidean distance, Mahalanobis distance and Kullback-Leibler divergence. They extended the test stimuli to cases of polysyllabic words in natural sentences which were synthesized by a state-of-the art unit selection TTS system: *rVoice* from Rhetorical Systems Ltd. They showed that none of the measures performed well in all cases and further research is needed to develop new distance measures so as to reliably predict audible discontinuities.

In an other study of Vepa and King [33], [34], linear dynamical model (Kalman filter) on LSF trajectories has been used for the computation of joint cost in unit selection speech synthesis. The model, after training, could be used to measure how well concatenated speech segments join together. The objective joint cost is based on the error between model prediction and actual observations. Linear dynamical model was used also for smoothing the LSF coefficients reducing the audible discontinuities. An advantage of this method is that the degree and extent of the smoothing is controlled by the model parameters which are learned from natural speech.

A novel discontinuity measure which accounts for both interframe incoherence and discrepancies in formant frequences/bandwidths was introduced by Bellegarda [22]. His metric is derived

through a pitch synchronous singular value decomposition (SVD) of the signal. He constructs a matrix with frames of speech in the vicinity of the concatenation point and with the use of SVD projects new frames to the space obtained from SVD. This alternative transform framework preserves those properties of the waveform which are globally relevant in the region of the concatenation. He compared his metric with Euclidean distance on MFCCs and found that listeners preferred the sentences synthesized with the new measure than the sentences synthesized with MFCCs.

Blouin and colleagues presented a joint cost function based on phonetic and prosodic features [35], [36]. This function is defined as a weighted sum of subcosts, each of which is a function of various symbolic and prosodic parameters. Weights were optimized using a multiple linear regression as a function of an acoustic measure of concatenation quality. This acoustic measure is calculated as a KL divergence on normalized LPC power spectra. Perceptual evaluation results indicated that the concatenation subcost weights determined automatically were better than hand-tuned weights, with or without applying F0 and energy smoothing after unit concatenation.

Kawai and Tsuzaki [37] compared acoustic measures and phonetic features in their ability to predict audible discontinuities. The acoustic measures were derived from MFCCs, mainly Euclidean distances between MFCCs of certain frames. A perceptual experiment was used to measure the degradation in naturalness due to signal discontinuities. Then, models were built to predict the degradation scores from the acoustic measures and phonetic features. The models used were multiple regression model; decision tree; neural network. The multiple regression coefficients were calculated under open and closed conditions of modeling and for acoustic measures and/or phonetic features. Phonetic features were found to be more efficient than acoustic measures in predicting audible discontinuities.

Syrdal and Conkie [38], [39] conducted studies to detect mid-vowel and mid-consonant concatenation discontinuities. They trained and tested several models, such as linear regression and classification and regression trees (CART), for predicting audible discontinuities. They used not

only acoustic features but also phonetic ones. Their results indicate that Euclidean cepstral distances were superior as acoustic features. Moreover, using phonetic features in CART models, the accuracy of prediction of concatenation discontinuities can significantly be improved.

To sum up, if there is a single conclusion that can be drawn from the above results, it is that no single concatenation 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 features that were used did not capture the phenomenon of audible discontinuity. Moreover, it is reasonable to say that the use of a perceptually motivated, nonlinear frequency scale is a good idea. Finally, phonetic features performed quite well, even so, the concatenation cost cannot be based solely on phonetic features.

# Chapter 3

# Feature Sets

In order to compute concatenation cost function, robust feature representation of speech signal has to be extracted. In this Chapter, five different feature representations are discussed in detail. Two of them have been used for the detection of discontinuities before, while the other three feature representations have been introduced by us.

## 3.1 LSF

A feature representation of a speech magnitude spectrum is that of Line Spectral Frequencies (LSF) [40]. LSF parameters have both well-behaved dynamic range and filter stability preservation property, therefore, they are very efficient in coding the LPC parameters. The computation of LSF parameters is based on LPC analysis.

For a given order $M$, LPC analysis results in a inverse filter

$$A_M(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + ... + a_M z^{-M} \tag{3.1}$$

which minimizes the residual energy [41].

We can extend the order of $A_M(z)$ to $(M + 1)$ without introducing any new information by

letting the $(M+1)^{th}$ reflection coefficient, $k_{M+1}$, be equal to 1 or $-1$. This is equivalent to setting the corresponding acoustic tube model completely closed or completely open at the $(M+1)^{th}$ stage. Thus, we have

$$P(z) = A_M(z) + z^{-(M+1)} A_M(z^{-1}) \tag{3.2}$$

when $k_{M+1} = 1$ (or, $(M+1)^{th}$ tube is completely closed), and

$$Q(z) = A_M(z) - z^{-(M+1)} A_M(z^{-1}) \tag{3.3}$$

when $k_{M+1} = -1$ (or, $(M+1)^{th}$ tube is completely open).

Zeros of $P(z)$ and $Q(z)$ constitute the LSF parameters. The LSF representation is rather artificial, however, it has very useful properties. The important properties of $P(z)$ and $Q(z)$ are:

- All zeros of $P(z)$ and $Q(z)$ are on the unit circle (Figure 3.1), so, only angle —which represents the frequency— of the zero is necessary;
- Zeros of $P(z)$ and $Q(z)$ are interlaced with each other; and
- Minimum phase property of $A_M(z)$ is easily preserved after quantization of the zeros of $P(z)$ and $Q(z)$, which is useful for speech coding.

The number of LSF is equal to $M$ which is the order of inverse filter. This number usually depends on the sampling frequency of the signal.

## 3.2 MFCC

The mel-cepstrum, introduced by Davis and Mermelstein [42], exploits auditory principles, as well as the decorrelating property of the cepstrum. In addition, the mel-cepstrum has proven to be one most successful feature representations in speech-related recognition tasks. Their dominance

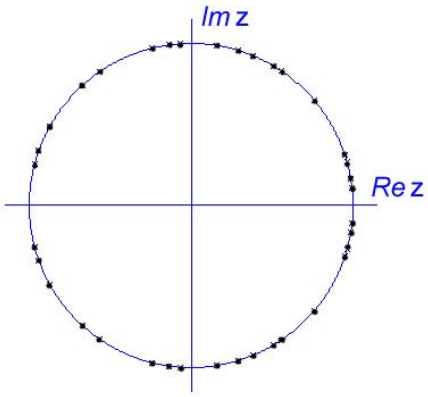



Figure 3.1: Zeros of $P(z)$ and $Q(z)$ are illustrated. All zeros are on the unit circle and each zero has its conjugate.

in speech recognition as well as in speaker identification/verification systems stems from their ability to represent the amplitude spectrum in a compact form. Apart from this, MFCC have been extensively used for the detection of audible discontinuities [17], [19], [21]. The mel-cepstum is computed as illustrated in Figure 3.2.

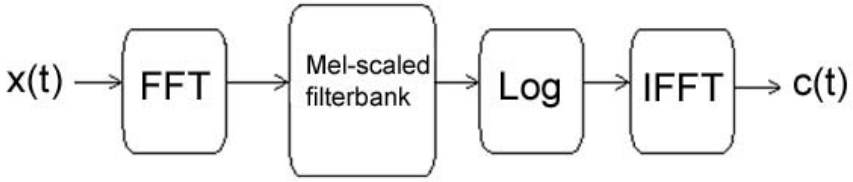



Figure 3.2: Frow diagram for the computation of MFCC.

Firstly, the evaluation of the magnitude spectrum, $|X(\omega)|$, through Short Time Fourier Transform is performed. Hamming window of 20 ms was used in short time Fourier analysis. Secondly, $|X(\omega)|$ is weighted by a series of filter frequency responses whose centers frequencies and bandwidths roughly match those of the auditory critical band filters. These filters follows a linear scale for the low frequencies (up to $1kHz$), and logarithmically increase for the high frequencies. In Figure 3.3, mel-scaled filter bank is illustrated. Then, the logarithm of the output of the filter is evaluated, and finally, inverse discrete Fourier transform (or, inverse discrete cosine transform)

is applied. The result of this procedure is the Mel-scaled frequency cepstral coefficients (MFCC). Similar to LSF the number of MFCC depends on the sampling frequency of the analyzed signal

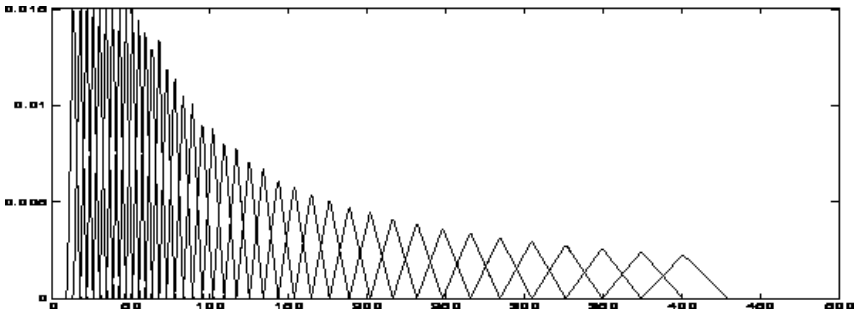


Figure 3.3: Filters of the filterbank. They are defined from the mel-scaled frequencies.

## 3.3 Nonlinear Harmonic Model

The first technique for analyzing speech signals is through a nonlinear harmonic model [23]. The model assumes the speech signal to be composed as a periodic signal, h[n], which is designated as sums of harmonically related sinusoids

$$h[n] = \sum_{k=-L(n_i)}^{L(n_i)} A_k[n] e^{j2\pi k f_0(n_i)(n-n_i)} \tag{3.4}$$

where $L(n_i)$ denotes the number of harmonics at $n = n_i$, $f_0(n_i)$ denotes the fundamental frequency at n = $n_i$, while $A_k$[n] is the time-varying complex amplitude,

$$A_k[n] = a_k(n_i) + (n - n_i) b_k(n_i) \tag{3.5}$$

where $a_k(n_i)$ and $b_k(n_i)$ are assumed to be complex numbers which denote the amplitude of the $k^{th}$ harmonic and the first derivative(slope) respectively.

The unknown complex amplitudes (eq. (3.5)) are estimated by minimizing a weighted time-

domain least-squares criterion with respect to $a_k(n_i)$ and $b_k(n_i)$,

$$\epsilon = \sum_{n=n_i-T_0}^{n=n_i+T_0} w^2[n](s[n] - h[n])^2 \tag{3.6}$$

where s[n] denotes the original speech signal, h[n] denotes the harmonic signal to estimate, w[n] denotes the weighted window (which is typically a Hanning window) and $T_0$ denotes the local fundamental period ($f_s/f_0(n_i)$), in samples. As eq. 3.6 indicates, the size of analysis window is two pitch periods

The estimation of the parameters is done by the steps presented below. Firstly, the local fundamental frequency, $f_0(n_i)$, is evaluated from the autocorrelation function of the speech signal around the analysis point. After this, in order to consider the whole spectrum, the number of harmonics, $L(n_i)$, is computed by $L(n_i) = \lfloor \frac{f_s}{2f_0(n_i)} \rfloor$ where $f_s$ denotes the sampling frequency and $\lfloor \rfloor$ denotes the floor operator. It is valid to consider the whole spectrum since the phonemes we tested are vowels, so, the noise part may be considered to be absent.

Solving the least-squares problem [23], we obtain a solution in closed form which is given by:

$$[a_1 a_2 ... a_L a_1^* a_2^* ... a_L^* b_1 b_2 ... b_L b_1^* b_2^* ... b_L^*] = (\mathbf{B^h W^h W B})^{-1} \mathbf{B^h W^h W s} \tag{3.7}$$

where "$h$" means Hermitian matrix operator, $\mathbf{W}$ is a $(2T_0 + 1) \times (2T_0 + 1)$ diagonal matrix with diagonal elements the values of the Hamming window while $L$ is abbreviation of $L(n_i)$. $\mathbf{B}$ is the concatenation of four matrices (i. e. $\mathbf{B} = [\mathbf{B_1}|\mathbf{B_2}|\mathbf{B_3}|\mathbf{B_4}]$) with elements

$$\begin{aligned}
(\mathbf{B_1})_{ml} &= E^{(m-T_0)(l+1)} \\
(\mathbf{B_2})_{ml} &= E^{-(m-T_0/2)(l+1)} \\
(\mathbf{B_3})_{ml} &= (m - T_0/2)(\mathbf{B_1})_{ml} \\
(\mathbf{B_4})_{ml} &= (m - T_0)(\mathbf{B_2})_{ml}
\end{aligned} \tag{3.8}$$

where $E = exp(2\pi j f_0(n_i))$, $m$ ranges from 0 to $2T_0$ and $l$ ranges from 0 to $L-1$. Consequently, $\mathbf{B}$ is a $(2T_0+1) \times (4L)$ matrix.

It is noteworthy that synthetic speech produced by this nonlinear harmonic model is indistinguishable from the original speech (modeling error is about -25dB). In addition, this model is not sensitive to small errors in pitch estimation.

## 3.4 AM & FM Decomposition

Teager [43], [44], in his work on nonlinear modeling of speech production, used the nonlinear operator

$$\Psi\{x[n]\} = x^2[n] - x[n-1]x[n+1] \tag{3.9}$$

on speech signals $x[n]$. This operator, also known as Teager-Kaiser energy operator, was used for the evaluation of the "energy" of a single-component signal [45], [46]. With energy we mean the generation energy which is analogous not only with the amplitude but also with the frequency. In Figure 3.4, the energy of an exponentially decaying signal with sinusoidal instantaneous frequency is shown.

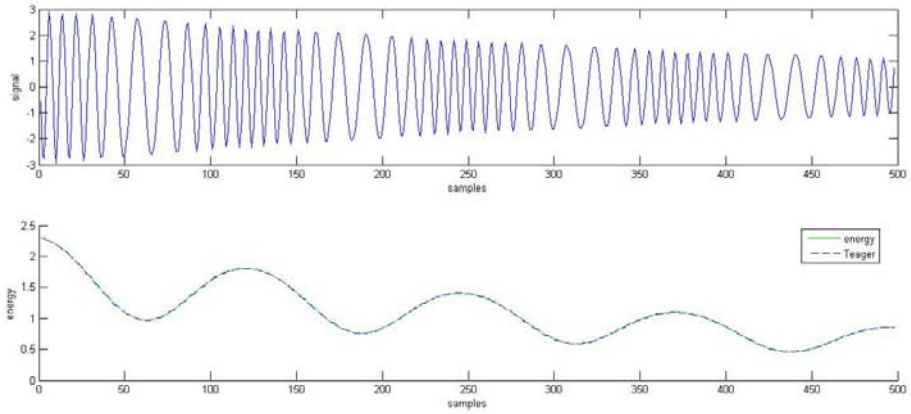



Figure 3.4: Energy of an AM&FM modulation signal. We consider as instant energy the product of instant amplitude and instant frequency. Observe the peaks of energy when signal oscillates faster and the valleys when signal oscillates slowly.

Teager energy operator was used by Maragos et al. [24] for the separation of amplitude from frequency modulations of an AM-FM signal. The separation of AM component from FM component is done through the Discrete Energy Separation Algorithm(DESA). The core of DESA are the following equations:

$$G[n] = 1 - \frac{\Psi\{y[n]\} + \Psi\{y[n+1]\}}{4\Psi\{x[n]\}} \tag{3.10}$$

$$\Omega[n] \approx \arccos(G[n]) \tag{3.11}$$

$$|a[n]| \approx \sqrt{\frac{\Psi\{x[n]\}}{1 - G^2[n]}} \tag{3.12}$$

where $y[n] = x[n] - x[n-1]$, $\Omega[n]$ is the instantaneous frequency and $a[n]$ is the instantaneous amplitude.

Examples of an AM, an FM and an AM&FM signal and there components after applying DESA are shown in Figures 3.5, 3.6 and 3.7.

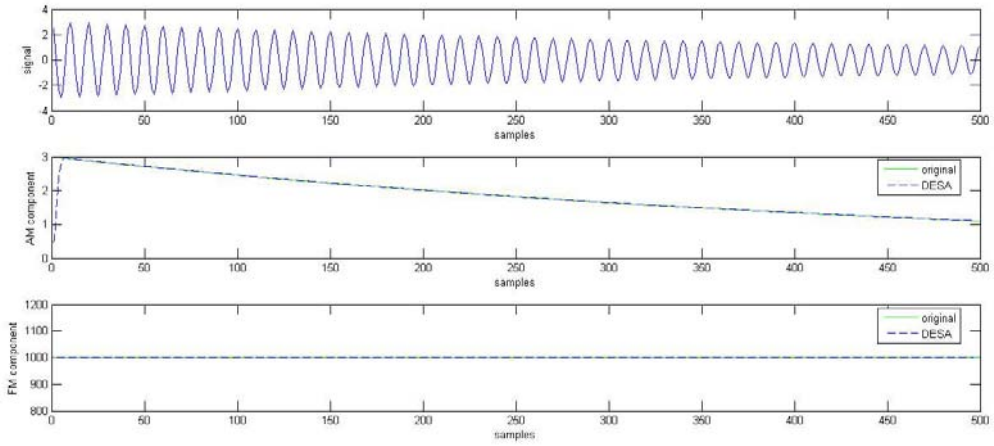



Figure 3.5: Separation into its components for an AM signal $x(t) = Ae^{-at}cos(2\pi f_r t)$

One application of DESA in speech analysis is the separation of a signal around a resonance in an amplitude and a frequency component [47]. The extraction of a single resonance is done by

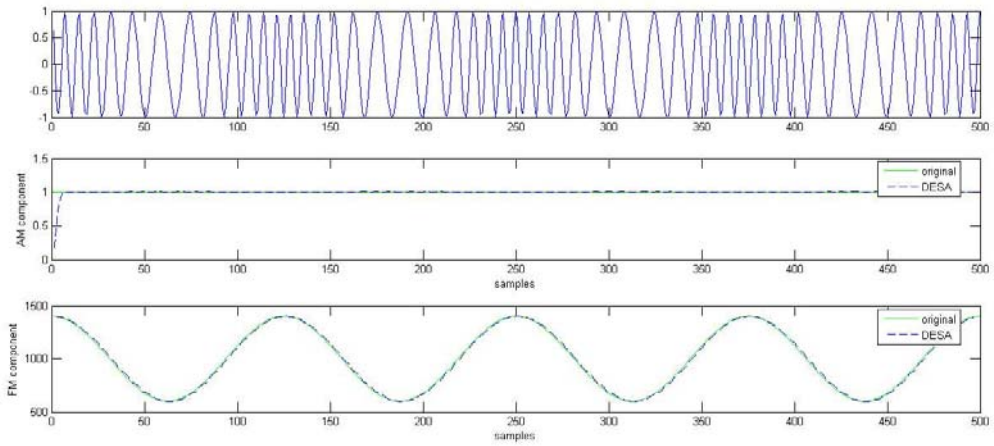



Figure 3.6: Separation into its components for an FM signal $x(t) = cos(2\pi f_r t + \beta sin(2\pi f_m t))$

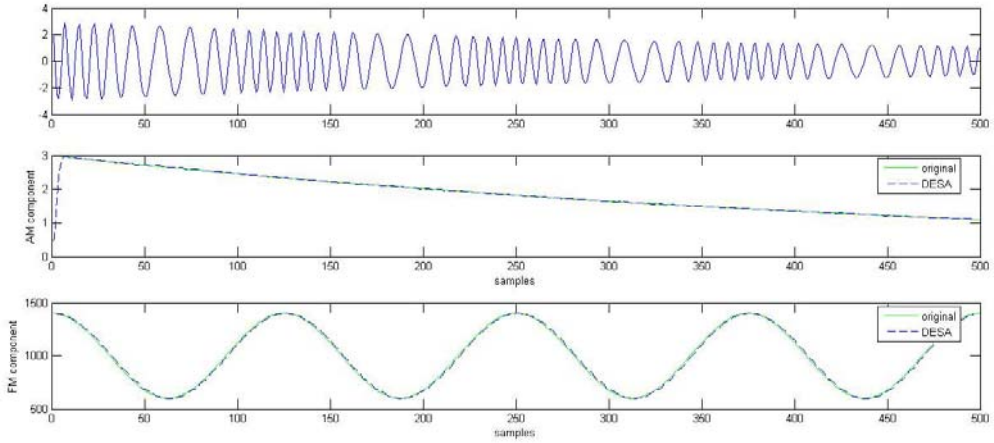



Figure 3.7: Separation into its components for an AM&FM signal $x(t) = Ae^{-at}cos(2\pi f_r t + \beta sin(2\pi f_m t))$

bandpass filtering the speech signal with a Gabor filter with impulse response defined by

$$h_G[n] = \exp(-a^2 n^2)\cos(\Omega_c n) \tag{3.13}$$

where $a$ controls the bandwidth of the filter and $\Omega_c$ is the central frequency of the resonance.

In our case, we decided to construct a filterbank of twenty Gabor filters. In our filter design the value of $a$ was selected to be 250, hence the bandwidth of each filter was approximately 425Hz. The central frequencies of the filterbank are equal spaced from 250Hz up to 5000Hz. The

filters of the filterbank are shown in Figure 3.8. The size of analysis window was 300 samples (approximately 20msec). Hence, twenty AM and FM components are extracted for each analyzed signal.

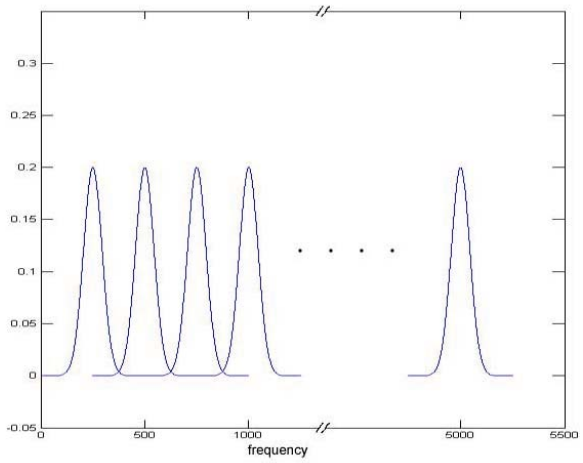



Figure 3.8: Filterbank of Gabor filters. These filters are equally spaced in frequency and they are slightly overlapping.

## 3.5 Auditory Models

The auditory system of humans consists of various parts that interact converting the sound pressure waves entering the outer ear into neural stimulus. Nowadays, it is possible to describe how signals are elaborated by the auditory system [48], but it is also possible to analyze signals using mathematical models that reproduce the auditory features [49], [50]. In this way, we have the possibility to understand which kind of representations —our higher levels— in the brain are used to isolate signals from noise, or to separate signals which have different pitches.

If we want to reproduce the same operations, we have to be able to work on representations similar to those used by our brain. In practice, this can be done using a mathematical auditory model, by which we analyze signals and then, depending on the application, extract useful information. In this thesis, Lyon's cochlea model is used for the detection of audible discontinuities. Regarding the software implementation of Lyon's auditory model, whose this section refers to, it

is that resulting from M. Slaney's work [51].

Lyon's auditory model [49], schematically illustrated in Figure 3.9, describes with particular attention the behavior of the cochlea, the most important part of the inner ear, that act substantially as a non-linear filter bank. Due to the variability of its stiffness, different places along the cochlea are sensible to sounds with different spectral content. In particular, at the base the cochlea is stiff, while going on, it becomes less rigid and more sensible to low frequency signals. This behavior is simulated in the model, by a cascade filter bank. The bigger the number of these filter the more accurate is the model. In front of these stages there is another stage that simulate the effects of the outer and middle ear (pre-emphasis). The number of the filters mainly depends on the sampling rate of the signals.

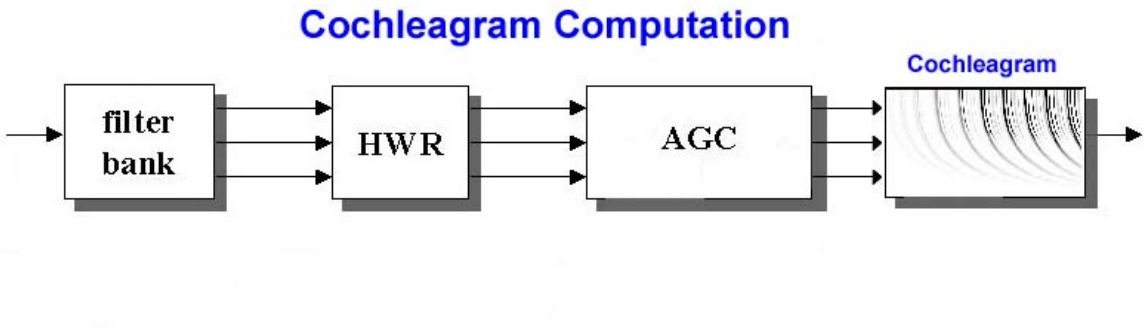



Figure 3.9: Lyon's auditory model for the inner ear (cochlea).

The next part of the model consists of an ideal half wave rectification (HWR), composed of a bank of HWRs which have the function to drop the negative portions of the waveform. They model the directional behavior of the inner hair cells, thus, they reduce the energy of the signal.

The final part of the model describes the adaptive features which work in our auditory system. This part consists of four automatic gain control (AGC) stages that are cascaded. The signals of each channel coming out of the HWR stages, pass through these four AGC stages. The value of the gain of each stage depends on a time constant, on the value of the preceding output sample and on the values of the preceding output samples of the adjacent channels. In this way it is possible to reproduce the masking effects. The different time constants simulate the different

adaptive times of our auditory system: the first AGC stage has the biggest time constant so that it reacts to the input signal more slowly, while the following stages have decreasing time constants. The outputs of these stages approximately represent the neural firing rates produced by the solicitation of various parts of the cochlea due to the sound pressure waves entering the outer ear.

The output of this stage is the cochleagram which is shown for a test signal in Figure 3.10. Cochleagrams look alike spectrograms in the sense that both are a two dimensional representation of a sound. In both representations X-axis represents time. However, Y-axis of cochleagram refers to neuron number of the ear and not frequency as in spectrogram. Moreover, the value of the cochleagram at each point refers to neural firing rate instead of energy.

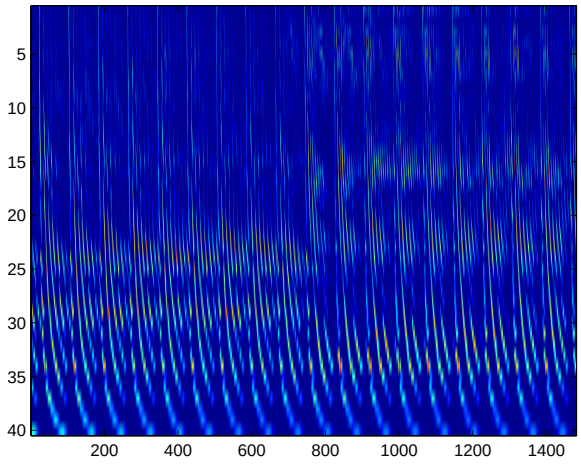



Figure 3.10: Cochleagram from a synthesized speech segment. In the middle of the segment, the concatenation point is evident.

# Chapter 4

# Discrimination of Features

To measure the difference between two feature vectors, a distance measure is needed. In this chapter, two metrics and three statistical approaches are presented. Usually, a distance measure satisfies the properties of symmetry, positive definiteness and efficient evaluation. However, this is not obligatory because we are interested in finding a cost which predicts audible discontinuities without any constraints.

## 4.1 Norms

The simplest distance measures are the $l_p$ norms. They satisfy the properties of symmetry, positive definiteness as well as triangular inequality. Assuming that $\mathbf{x}$ and $\mathbf{y}$ are vectors with $\mathbf{x} = [x_1,\ x_2,\ ...,\ x_d]^T$ and $\mathbf{y} = [y_1,\ y_2,\ ...,\ y_d]^T$, we have the following norms.

### 4.1.1 $l_1$ or Absolute Norm

$l_1$ or absolute distance is computed as the sum of the absolute difference of the parameters of two feature vectors.

$$D_{abs}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{d} |x_i - y_i| \tag{4.1}$$

### 4.1.2 $l_2$ or Euclidean Norm

$l_2$ also called Euclidean distance is the square root of the sum of the squares of the difference of the parameters of two feature vectors, or more easily,

$$D_{Eu}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2} \tag{4.2}$$

## 4.2 Fisher Linear Discriminant

Suppose that we have a set of N d-dimensional samples $\mathbf{x_1}$,...,$\mathbf{x_N}$, $N_0$ samples be in the subset $D_0$ and $N_1$ samples be in the subset $D_1$. If we form a linear combination of the elements of $\mathbf{x}$, we obtain the scalar dot product

$$y = \mathbf{w^T x} \tag{4.3}$$

and a corresponding set of N samples $y_1$,...,$y_N$ that is divided into the subsets $Y_0$ and $Y_1$. This is equivalent to form a hyperplane in d-space which is orthogonal to $\mathbf{w}$ (Figure 4.1).

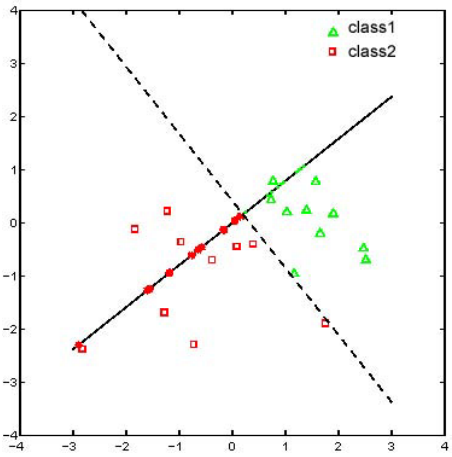



Figure 4.1: Example of Fisher's linear discriminant between two classes.

The direction of $\mathbf{w}$ is important for adequate separation [25] and is given by

$$\mathbf{w} = \mathbf{S_W^{-1}}(\mu_\mathbf{0} - \mu_\mathbf{1}) \tag{4.4}$$

where

$$\mathbf{S_W} = \sum_{\mathbf{i=0}}^{\mathbf{1}} \sum_{\mathbf{x}\in D_i} (\mathbf{x} - \mu_\mathbf{i})(\mathbf{x} - \mu_\mathbf{i})^\mathbf{T} \tag{4.5}$$

is the scatter matrix, and

$$\mu_\mathbf{i} = \frac{1}{\mathrm{N_i}} \sum_{\mathbf{x}\in D_i} \mathbf{x} \ , \qquad \mathrm{i} = 0, 1. \tag{4.6}$$

are the mean values of the two distributions.

Since Fisher's linear discriminant (FLD) projects feature vectors to a line it can also be viewed as an operator which is defined by

$$FLD\{\mathbf{x}\} = \sum_{i=1}^{d} w_i x_i \tag{4.7}$$

where $w_i$ are the elements of $\mathbf{w}$. If $x_i$ are real positive numbers, this is a kind of weighted version of $l_1$ norm (yet, in this case weights can take negative values).

## 4.3 Linear Regression

In statistics, linear regression (LR) is a method that attempts to model the relation between two variables by fitting a linear equation to observed data (Figure 4.2). One variable is considered to be an explanatory or input variable, and the other is considered to be a dependent or output variable.

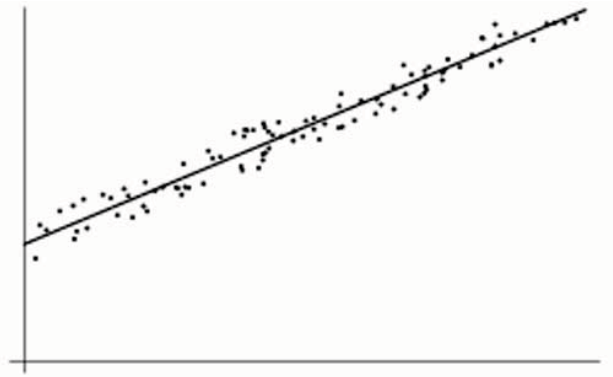


Figure 4.2: Example of linear regression fit.

A linear regression model is typically stated in the form [52], [53]

$$y_i = \alpha^t \mathbf{x_i} + \beta + \epsilon \tag{4.8}$$

where $\mathbf{x_i}$ is the input value which has d dimensions, $y_i$ is the observed value, $\alpha$ and $\beta$ are the parameters of the model and finally, $\epsilon$ is the unpredicted or unexpected variation in the response, conventionally referred to as error.

For the estimation of parameters $\alpha$ and $\beta$, least-squares method is used. In other words, minimization of the square of all residual errors is performed. Assume that we have $N$ input/output pairs $(\mathbf{x_i}, y_i)$. Then, in matrix form we obtain

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} \mathbf{x_1}^t & 1 \\ \vdots & \vdots \\ \mathbf{x_N}^t & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_N \end{bmatrix} \tag{4.9}$$

or, in compact form

$$\mathbf{Y} = \mathbf{X} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \epsilon \tag{4.10}$$

The optimum solution in least-squares sense is given by:

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{Y} \tag{4.11}$$

Actually, $\beta$ is not important, because we are interested only in the direction (or slope) of the model which is given by $\alpha$.

Finally, linear regression is similar to Fisher's linear discriminant since both methods are linear and optimal for normal distributions. However, their parameters are estimated by different ways. Linear regression through least-squares method while Fisher's linear discriminant through Bayes

classification.

## 4.4 Quadratic Discriminant (QD)

Fisher's linear discriminant is the best solution in Bayesian sense, when samples of the two classes follow normal distribution with equal covariance matrices. However, if covariance matrices are different, the best solution is that of quadratic form [25], assuming again a normal distribution. Thus, following as much as possible the symbolism of Section 4.2 we have:

$$\mathbf{x_i} \in D_0 \sim N(\mu_\mathbf{0}, \mathbf{\Sigma_0}) \tag{4.12}$$

and

$$\mathbf{x_i} \in D_1 \sim N(\mu_\mathbf{1}, \mathbf{\Sigma_1}) \tag{4.13}$$

with $\mathbf{\Sigma_0} \neq \mathbf{\Sigma_1}$.

Then, Bayes classifier (or discriminant) take the form:

— choose class 0 when $g_0(\mathbf{x}) > g_1(\mathbf{x})$ and

— choose class 1 when $g_0(\mathbf{x}) < g_1(\mathbf{x})$

where

$$g_i(\mathbf{x}) = \mathbf{x}^t\mathbf{W_i}\mathbf{x} + \mathbf{w_i}^t\mathbf{x} + w_{0,i} \qquad i = 0 \; or \; 1. \tag{4.14}$$

and

$$\mathbf{W_i} = -\frac{1}{2}\mathbf{\Sigma_i}^{-1} \qquad and \qquad \mathbf{w_i} = \mathbf{\Sigma_i}^{-1}\mu_\mathbf{i} \tag{4.15}$$

The constant $w_{0,i}$ is not necessary to be computed since we adjust it through false alarm as we will see in next the chapter. Finally, $\mathbf{\Sigma_i}$ and $\mu_\mathbf{i}$ are evaluated from the samples as,

$$\mu_\mathbf{i} = \frac{1}{\mathrm{N_i}} \sum_{\mathbf{x} \in D_i} \mathbf{x} \; , \qquad \mathrm{i} = 0, 1. \tag{4.16}$$

and

$$\mathbf{\Sigma_i} = \frac{1}{\mathrm{N_i}} \sum_{\mathbf{x} \in D_i} (\mathbf{x} - \mu_\mathbf{i})(\mathbf{x} - \mu_\mathbf{i})^\mathbf{t} \ , \qquad \mathrm{i} = 0, 1. \tag{4.17}$$

# Chapter 5

# Databases

One approach to comparing alternative concatenation cost functions is to implement each function in a synthesizer and compare the synthesized speech of the two systems. This is time consuming and requires repeated perceptual listening tests each time the concatenation cost function is altered. We use a methodology that requires only a single set of listening tests but still allows objective comparisons to be made between alternative join cost formulations. In this method, synthetic speech stimuli are generated in which there are a range of joint qualities. The stimuli are then rated by listeners. Comparison of join cost functions is then achieved by computing the correlations and detection rates between concatenation costs and listener ratings. Good join costs correlate strongly with listener ratings of perceptual join discontinuity.

Two distinct databases were used for the comparison of the features and methods described in previous chapters. First database was constructed by Klabbers and Veldhuis [13] at Holland, while the second was constructed by Stylianou and Syrdal [19] at AT&T Labs. Similarities and differences of the two databases are also discussed.

## 5.1 Database No.1

For the construction of the database, a speech synthesizer is needed. IPOs speech-synthesis system Calipso which employs diphones as concatenative units from a professional female speaker was used. Diphones have been excised from nonsense words. For instance, consonant–vowel (CV) and vowel–consonant (VC) diphones are excised from nonsense words of the form $C@CVC@$. In order to reduce the data set to manageable proportions, this study was restricted to three Dutch vowels in this database, i.e., the vowels /a:/, /i/, /u/ (in SAMPA notation). These vowels were chosen because they cover the extremes in the vowel space.

The stimuli consisted of concatenated left CV and right VC diphones, which were excised from the nonsense words $C_l@C_lVC_l@$ and $C_r@C_rVC_r@$. The stimuli consisted of three vowel conditions in the context of all consonant pairs that can occur in C and C position. Therefore, the total number of stimuli is $23 \times 3 \times 21 = 1449$ CVC triples. So, for instance, the diphones /du/ and /uk/ that form the stimulus /duk/ were extracted from the nonsense words d@dud@ and k@kuk@. The diphones were created using a variation of sinusoidal analysis/synthesis technique. No spectral smoothing was applied at the boundary. In the stimuli, the consonant portions were cut off to prevent them from influencing the perception of the diphone transition in the middle of the vowel. Moreover, fading was used to smooth the transition from silence to vowel and vice versa. Because all stimuli were presented in isolation, the stimulus duration had to be long enough to be able to perceive the transition at the diphone boundary. The duration of the vowels was fixed to 130 ms with the diphone boundary located exactly in the middle of the vowel. The signal power of the second diphone was scaled to match that of the first diphone.

Five participants with a background in psycho-acoustics or phonetics participated in the perceptual experiment. It was a within-subjects design meaning that each subject received all stimuli in random order. For each stimulus, the participants had to judge the transition at the diphone boundary as either smooth (0) or discontinuous (1). The experiment was divided into

three hourly sessions which were held on different days, with a short break halfway through each session. The session order was different for all participants. The experiment started with a familiarization phase in which two stimuli were presented for each vowel, one being clearly smooth and the other being clearly discontinuous. The setup of this experiment results in very critical observations because

(a) the vowels have been placed out of context and

(b) subjects are forced to make a binary decision. This provides a more critical test than when using real speech.

The participants found the task difficult, but felt they had been able to make consistent decisions after the familiarization phase. As a consistency check, they presented two stimuli, one clearly smooth, the other clearly discontinuous, ten times at random positions in the total stimulus list. All participants were 100% consistent in their scoring of these two stimuli. Between participants there was more variability, as some participants applied a stricter threshold than others. In order to reduce the variability between participants, a majority score was calculated, i.e., a stimulus was marked as discontinuous when three out of five listeners perceived it as such. Summing the majority scores obtained in the experiment for each of the vowels, we get the percentage of perceived discontinuities. The results show that the number of audible discontinuities is particularly high for /u/ and comparatively low for /a/. This is due to the fact /u/ has the greatest amount of coarticulation and the /i/ has the smallest amount, closely followed by /a/.

## 5.2 Database No.2

Second database used for our research was constructed at AT&T Labs. It is consisted of 2016 monosyllabic words which were generated by concatenative synthesis using an acoustic inventory of recordings from a native American female speaker. The sampling frequency of these recording was 16kHz. The context of the inventory contained 336 monosyllabic test words that constitute

the Modified Rhyme Test [54]. Synthetic words were obtained by simple concatenation of raw waveforms using each time two halves of original words. The concatenation point was approximately obtained in the middle of the vowel. In order to avoid linear phase mismatches between the concatenated parts, a cross correlation function was used. From listening tests we may say that, in general, pitch continuation was preserved. The 336 spoken words were separated into 56 groups of 6 words. Each group had words with same vowel nucleus but different initial or final consonant(s). Therefore, for each group 36 synthetic words (test stimuli) were constructed (all possible combinations of the 6 recorded words). These 36 synthetic words constitute a subtest. Every subtest contained 6 "synthesized" words which actually were human spoken words and we used them for validation purposes.

The listening task was conducted in a quiet office room using headphones. Listeners were presented with a test stimulus along with a decision in order to familiarize themselves with the listening test. After this training period, listeners started to hear the test words followed by a single interval of forced choice (Yes/No) depending on whether or not they had heard a concatenation discontinuity. The number of subtests listened by the participants was 386.

Twelve listeners participated in the perceptual test. Most of the participants had experience in listening to synthetic speech. As a validation check, we tested how many of the continuous words were considered as discontinuous. A subtest was rejected if more than one continuous word was considered as discontinuous. This way, 62 subtests were rejected from the database while 324 subtests remained.

Finally, two numbers were assigned to each test stimulus. First number counted how many listeners perceived test stimulus discontinuous while second number counted how many listeners perceived test stimulus continuous. A synthetic speech signal was considered discontinuous(or continuous) if the first number was greater(or less) to the second number. Rarely, when a tie occurred synthetic signal was considered as discontinuous.

## 5.3 Similarities and Differences

Due to the fact that both databases constructed for the same purpose it is reasonable to have many similarities. On the other hand these databases were constructed from different researchers, therefore there are also differences.

Firstly, in both databases special treatment were done for the elimination of phase mismatches. Phase mismatch is another cause of audible discontinuities, however, we are not interesting on such cases since it is easy to eliminate them. Furthermore, energy of the right unit was normalized with respect to the left unit in order to avoid energy jumps. In addition, both databases synthesize vowels and not consonants. Database from AT&T uses more vowels (possibly all the English vowels) than the database from Holland which uses only three vowels. Lastly, listeners in both cases had to make a binary and forced decision (continuous or not).

On the other hand, stimuli duration in the databases varies. In Klabber's database stimuli presented to the listeners have duration 130 ms only while in Stylianou's database stimuli were the whole synthesized word. What is more, using signal processing techniques Klabbers normalize the fundamental frequency in both units to 200Hz. Stylianou et al. have left the fundamental frequency untouched without any signal processing alternations. We are not sure which approach is better since both have advantages and disadvantages. Needless to say, the systems that were used for synthesizing the stimuli were different.

Finally, an issue which is important for valid results is the point were the evaluation of the features is done. In the database of AT&T analysis has been done at the concatenation point (Figure 5.2). In the database of Klabbers et al. analysis has been done at the edge of the concatenation point (Figure 5.1). It is important to make the analysis as close as possible to the concatenation point because we are trying to take advantage of dynamic information which may change rapidly within few pitch periods.

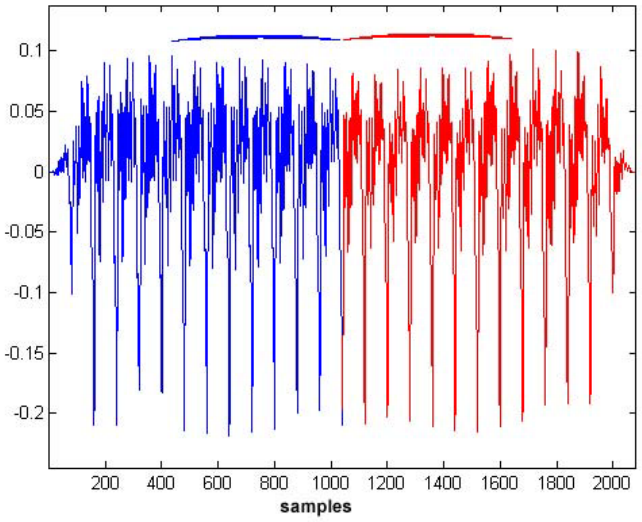



Figure 5.1: Where the analysis is done for the database of Klabbers et al.

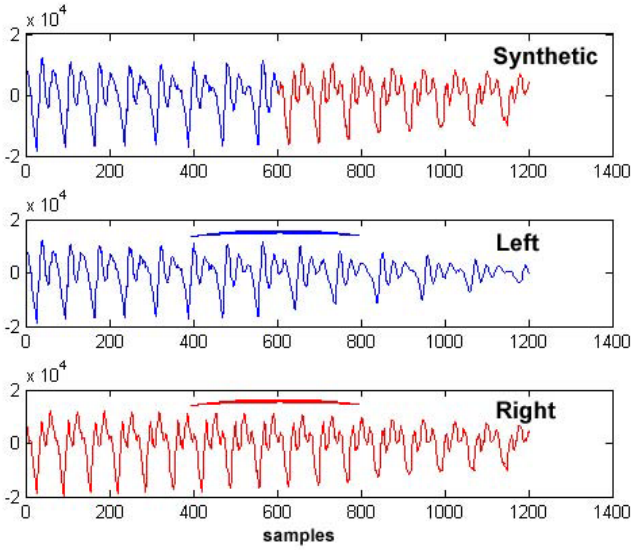



Figure 5.2: Where the analysis is done for the database of AT&T. First subplot shows the synthetic speech signal, while, the other two show the left and right speech segments.

# Chapter 6

# Results

## 6.1 ROC Curves

One way to evaluate the performance of the various distance measures is based on the detection rate, $P_D$, versus false alarm rate, $P_{FA}$ (Figure 6.1). This is done through receiver operator characteristic (ROC) curves [55], coming from signal detection theory. The procedure works as follows.

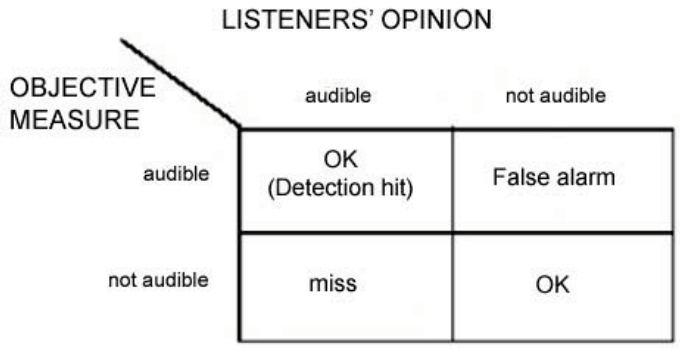



Figure 6.1: In a binary hypothesis test we must choose between two hypotheses. Depending on the choice there are four cases.

For each measure, $y$, two probability density functions, $p(y|0)$ and $p(y|1)$ were computed depending on the results from the perceptual test; if the synthetic word was perceived as continuous (0), and (1) if it was perceived as discontinuous by the listeners. In Figure 6.2, probability

functions for both continuous and discontinuous stimuli are shown. Then, the detection rate and false alarm for that measure, $y$, is computed as:

$$P_D(\gamma) = \int_{\gamma}^{\infty} p(y|1)\, dy \tag{6.1}$$

and

$$P_{FA}(\gamma) = \int_{\gamma}^{\infty} p(y|0)\, dy \tag{6.2}$$

where $\gamma$ defines the threshold (Figure 6.2).

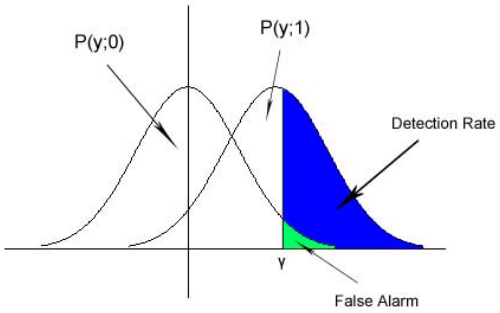



Figure 6.2: Probability density functions for continuous and discontinuous test stimuli.

A plot of pairs $(P_D(\gamma), P_{FA}(\gamma))$ for all values of $\gamma$ constitutes a receiver operating characteristic (ROC) curve. See Figure 6.3 for a schematic representation of ROC curves. The straight line represents the chance level meaning that a measure gives no information. The further the curve extends to the upper left corner, the better the measure serves as a predictor. This indicates that the two probability density functions and are moving away from each other, thus increasing the hit rate and decreasing the false alarm rate.

## 6.2 Correlation Coefficient

In probability theory and statistics, correlation coefficient measures the degree of correlation, adapted to the nature of data. To put it in another way, correlation coefficient indicates the

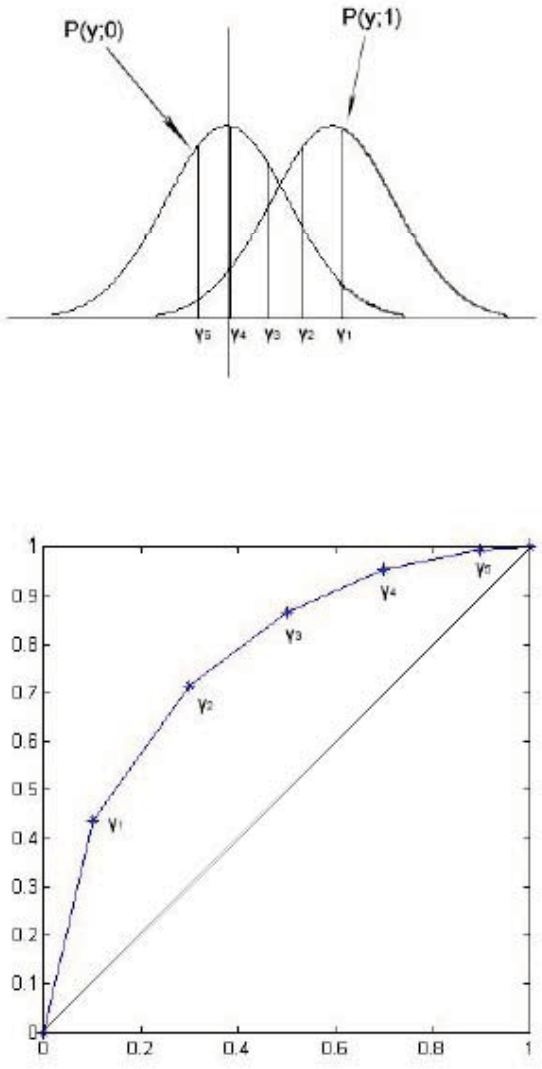



Figure 6.3: Receiver operator characteristic curve. $P_D$ and $P_{FA}$ are computed for various values of the threshold, $\gamma$.

strength and direction of a linear relationship between two random variables. In Figure 6.4, square of correlation coefficient is shown for various degrees of correlation between two random variables.

The correlation $\rho_{X,Y}$ between two random variables $X$ and $Y$ with expected values $\mu_X$ and $\mu_X$, also known as Pearson's product-moment, is defined as:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{E[(X-\mu_X)^2]E[(Y-\mu_Y)^2]}} \tag{6.3}$$

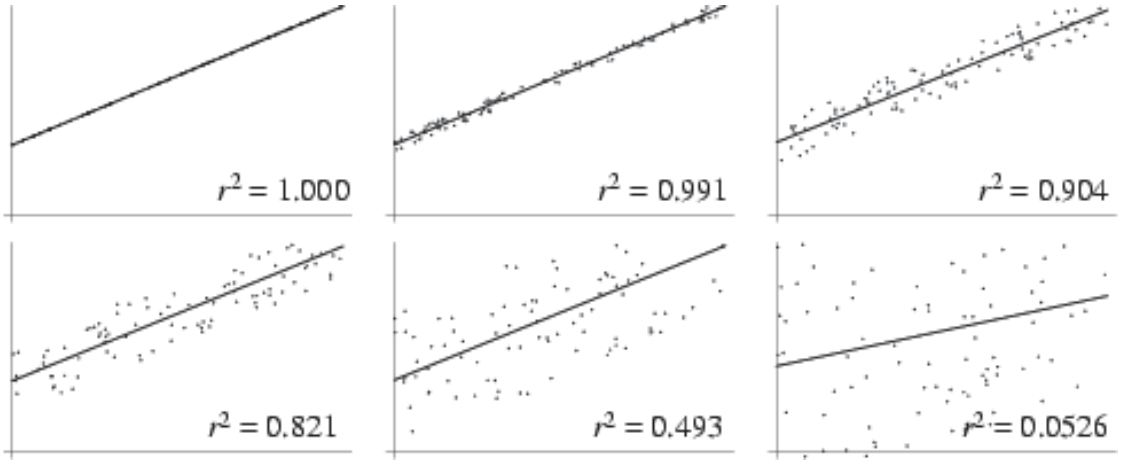



Figure 6.4: Correlation coefficient for various values of correlation.

Correlation coefficient takes values between $[-1, 1]$. Value 1 means that the correlation between the variables is strong while value 0 means that the variables are uncorrelated.

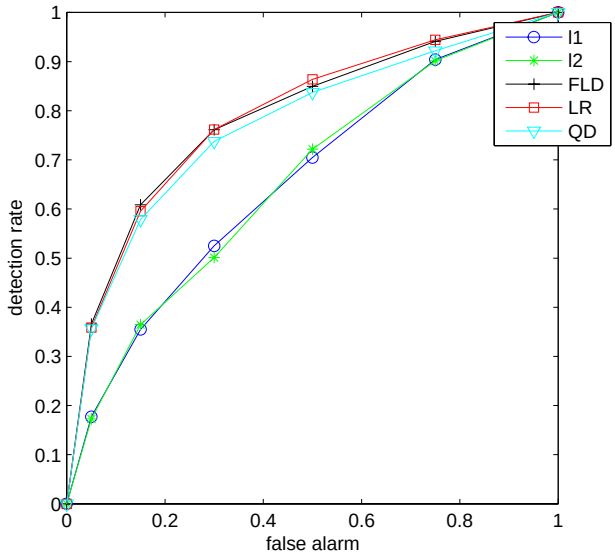
When we have N measurements of $X$, $Y$ pairs, written as $(x_1, y_1)$, ..., $(x_N, y_N)$, correlation coefficient is computed by:

$$r_{xy} = \frac{N \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{N \sum x_i^2 - (\sum x_i)^2} \sqrt{N \sum y_i^2 - (\sum y_i)^2}} \tag{6.4}$$

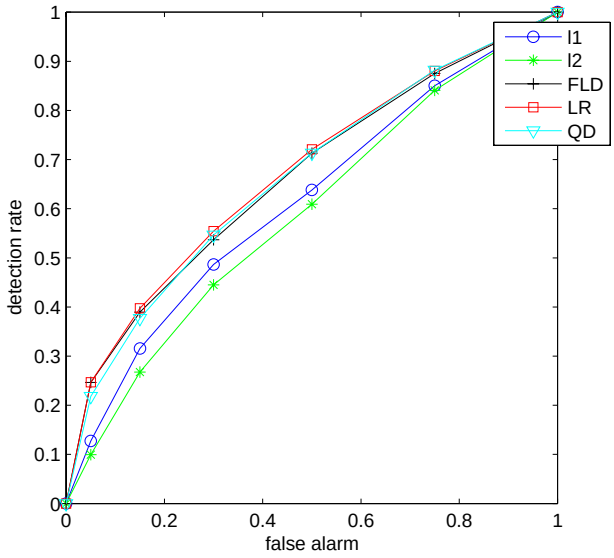
## 6.3 Results

In this section, ROC curves and tables with correlation coefficients are presented in detail. For the statistical methods such as Fisher linear discriminant and linear regression, the training was done on the 80% of the database, while the testing was done on the remaining 20% of the database. Furthermore, when a detection rate is given without referring false alarm, it is assumed that false alarm is 5%.

The number of LSF, as we have already said, primarily depends on the sampling frequency. Sampling frequency is 16 $kHz$ in both databases, therefore the number of LSF parameters is 18. LSF with FLD, LR or QD performed better than $l_1$ or $l_2$ in both database (Figure 6.5). In addition, LSF performed quite satisfactory in database of Klabbers et al. (detection rate 37%),




(a) Database No.1

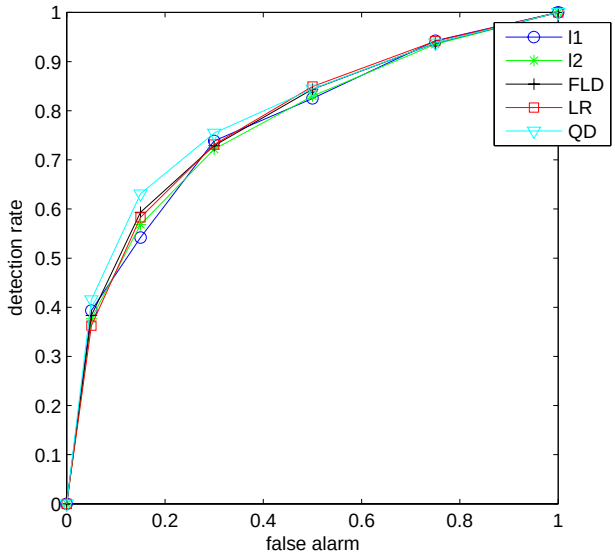



(b) Database No.2

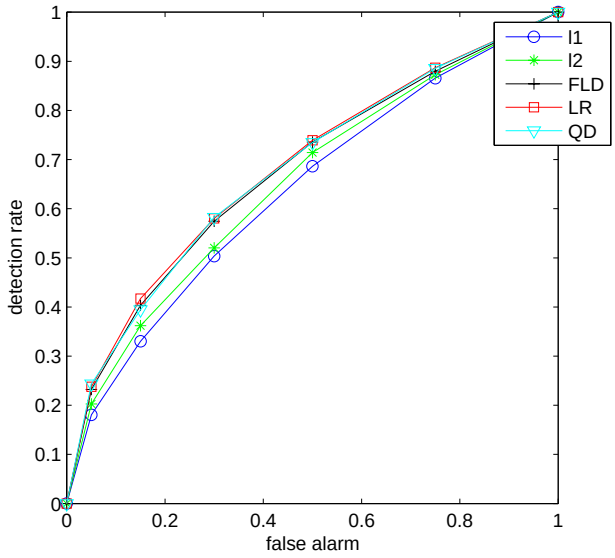
Figure 6.5: ROC curves for LSF and various discriminant functions are shown.

while in AT&T's database detection rate was 25%.

MFCC performed the same with all discriminant methods in database No.1 (Figure 6.6). In less extend, the same can be said for database No.2. This may result from the decorrelation property of MFCC parameters which results in equivalence of the coefficients. What is more,




(a) Database No.1

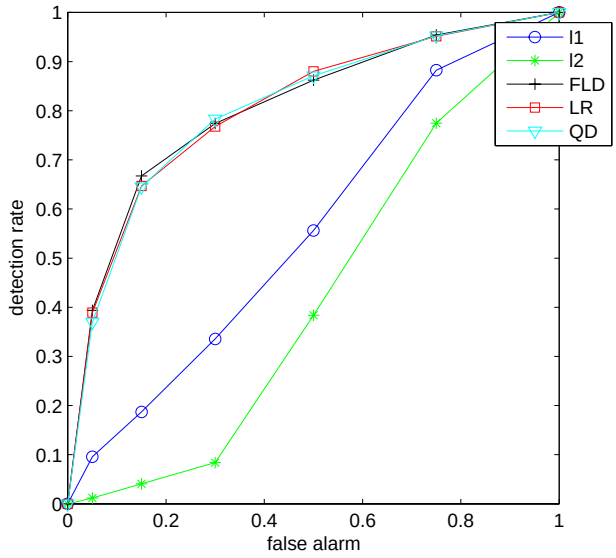



(b) Database No.2

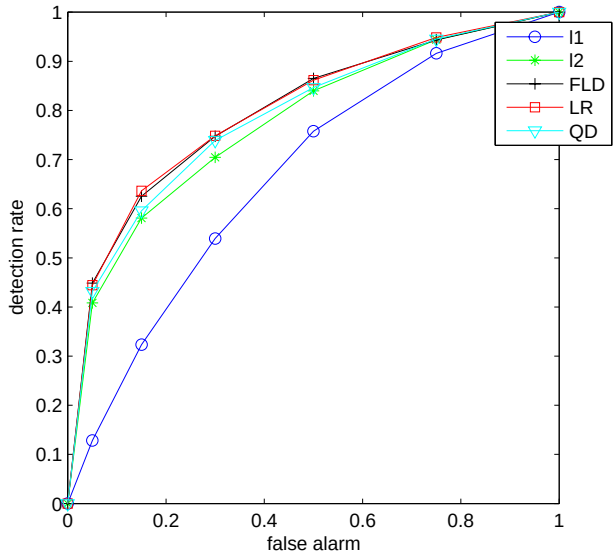
Figure 6.6: ROC curves for MFCC and various discriminant functions are shown.

MFCC performed better in database No.1 having detection rate of 40%, while in database No.2 detection rate was 24%. MFCC parameters are slightly better than LSF parameters.

Discrimination methods use vectors with real valued parameters (i.e. $\mathbf{x} \in \mathbb{R}^d$). However, our feature parameters except for LSF and MFCC are either complex numbers of functions



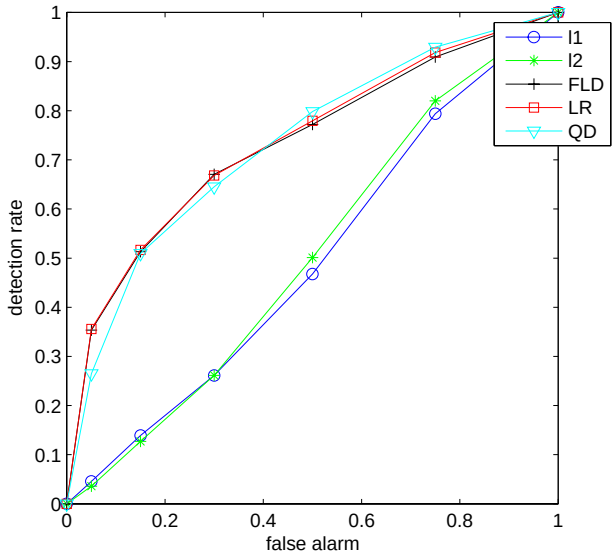

(a) Database No.1



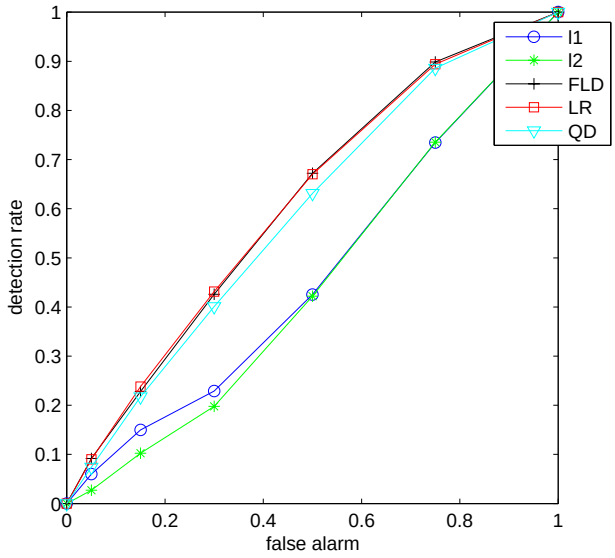

(b) Database No.2

Figure 6.7: ROC curves for amplitude of nonlinear harmonic model, $a_k$, and various discriminant functions are shown.

of time. Consequently, there is a need for converting them in real numbers. For the complex parameters of nonlinear harmonic model, there are two options; either take the absolute difference of the absolute value of the number, $||x_L| - |x_R||$, or take the absolute difference of the numbers, $|x_L - x_R|$. We used the latter, since, it incorporates not only amplitude but also phase information.




(a) Database No.1

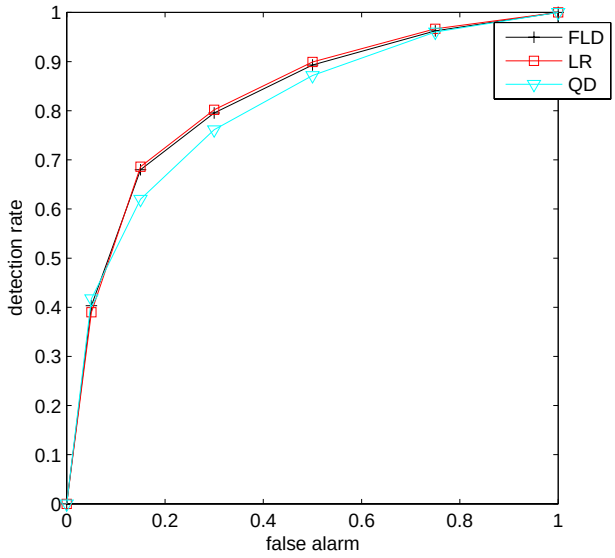



(b) Database No.2

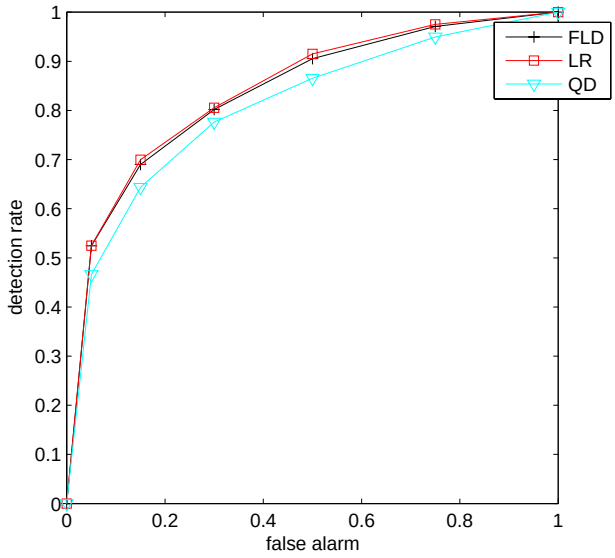
Figure 6.8: ROC curves for slope of nonlinear harmonic model, $b_k$, and various discriminant functions are shown.

Moreover, in order to keep the size of the measured vectors small while preserving the important information from a speech frame, we have decided to prune the size vector of complex amplitudes to the twenty first harmonics.

The amplitude of the nonlinear harmonic model, $a_k$, did not perform well for $l_1$ in both



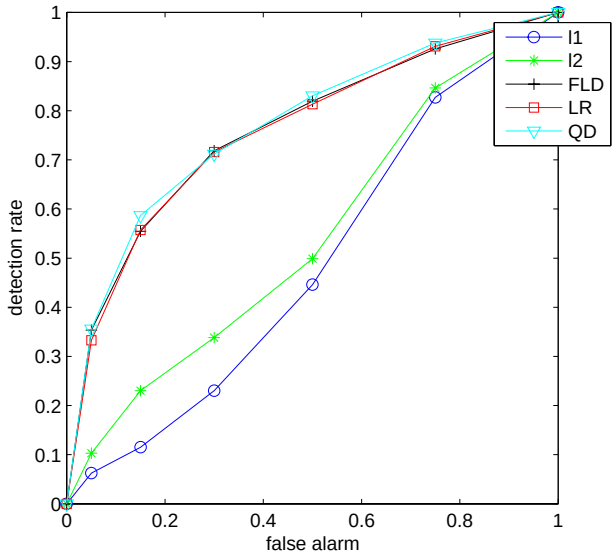

(a) Database No.1



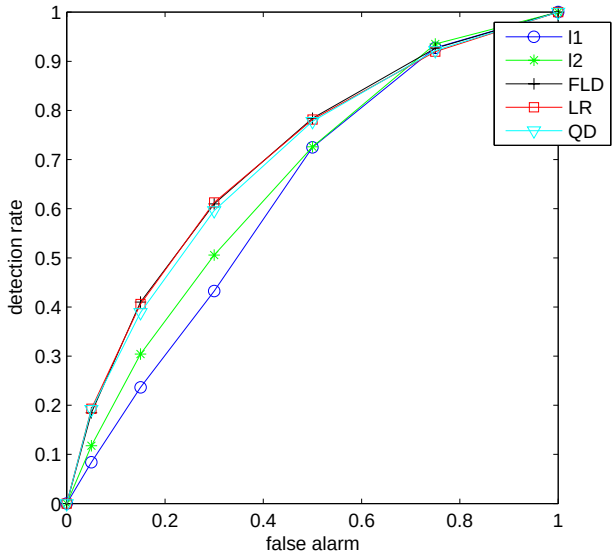

(b) Database No.2

Figure 6.9: ROC curves for both amplitude $a_k$, and slope $b_k$, of nonlinear harmonic model, and various discriminant functions are shown.

databases(Figure 6.7). Nevertheless, statistical methods provided much higher scores. Database No.1 gave detection rate 40% while database No.2 gave detection rate 45%. The slope of the nonlinear harmonic model, $b_k$, performed poor for AT&T's database for all the discriminant functions (Figure 6.8). However, in database of Klabbers' et al., it performed well giving detection



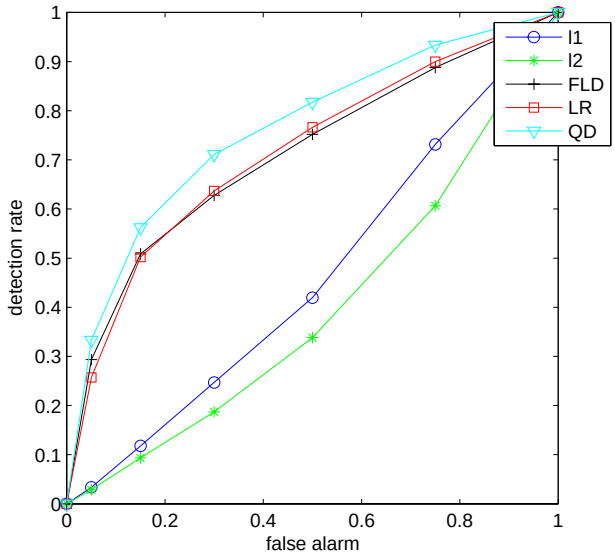

(a) Database No.1



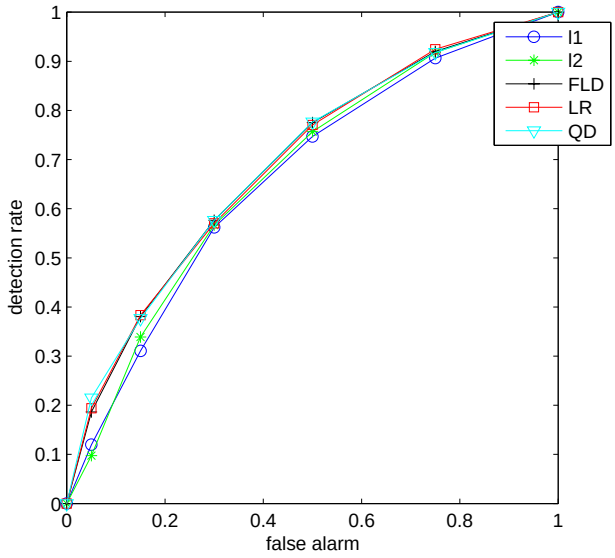

(b) Database No.2

Figure 6.10: ROC curves for AM components, and various discriminant functions are shown.

rate of 35% when FLD or LR is used. Only statistical methods can be used when $a_k$ and $b_k$ are combined by concatenation. Detection rates are high for all the discrimination methods in both databases (Figure 6.9). In database of Klabbers et al., detection rate was 41%, while in AT&T's database was detection rate 53%. This was by far the highest detection rate for AT&T's database.



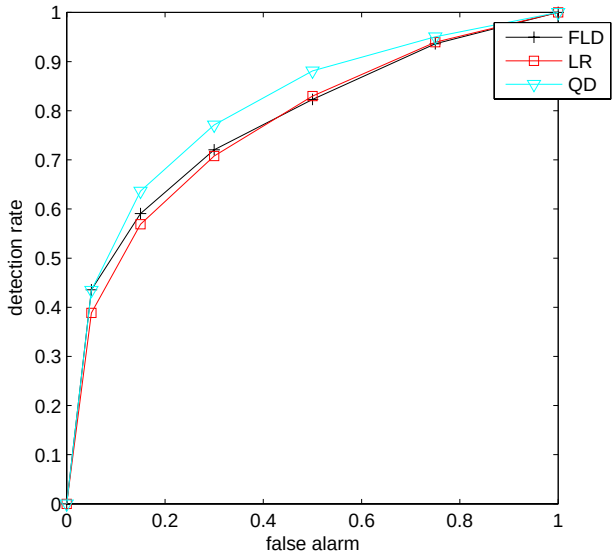

(a) Database No.1



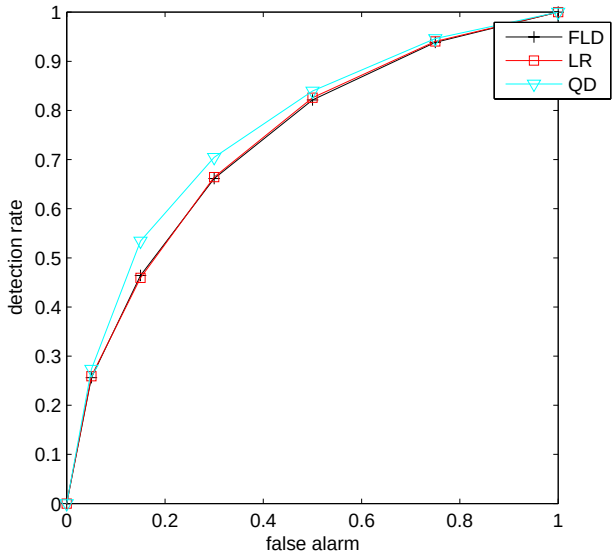

(b) Database No.2

Figure 6.11: ROC curves for FM components, and various discriminant functions are shown.

In the case of AM–FM decomposition, our features are signals. We decide to take the sum of the absolute difference between the left and the right signal as scalar features. Under this condition, it is necessary left and right (AM or FM) signals to be aligned, however, only in AT&T's database it was achievable. In database No.1, AM component performed well with



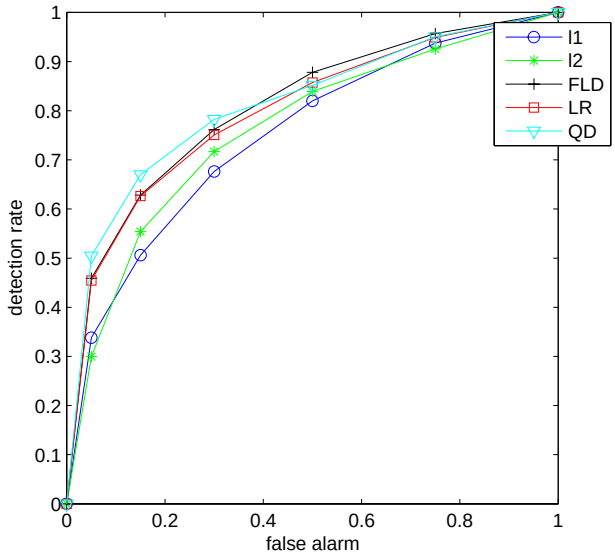

(a) Database No.1



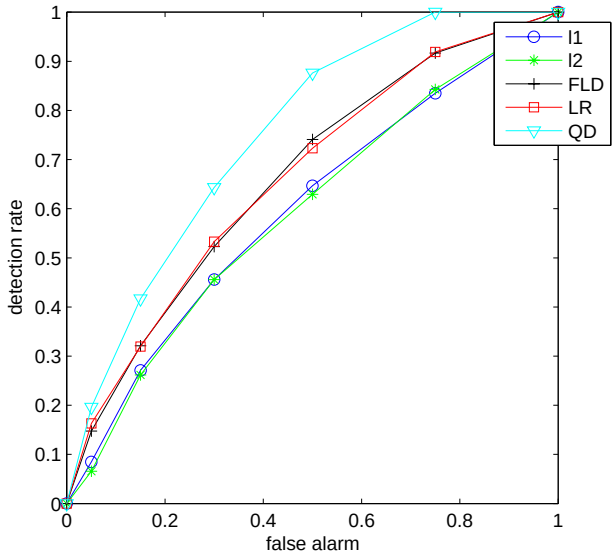

(b) Database No.2

Figure 6.12: ROC curves for both AM and FM components, and various discriminant functions are shown.

statistical disctiminant functions, but very poor with $l_1$ and $l_2$ where the scores were less than chance threshold (Figure 6.10). In database No.2 the detection rates where more robust, but not very high. Even though, $l_1$ and $l_2$ performed worse than FLD, LR ar QD. Similar results to AM component are obtained for the FM component (Figure 6.11). When the combination of AM




(a) Database No.1

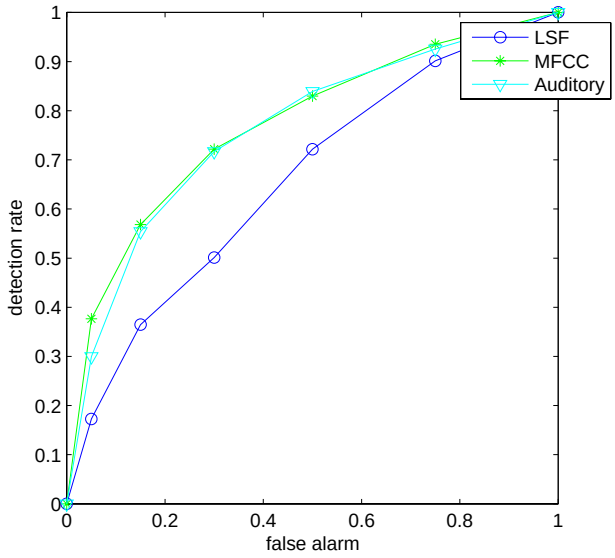



(b) Database No.2

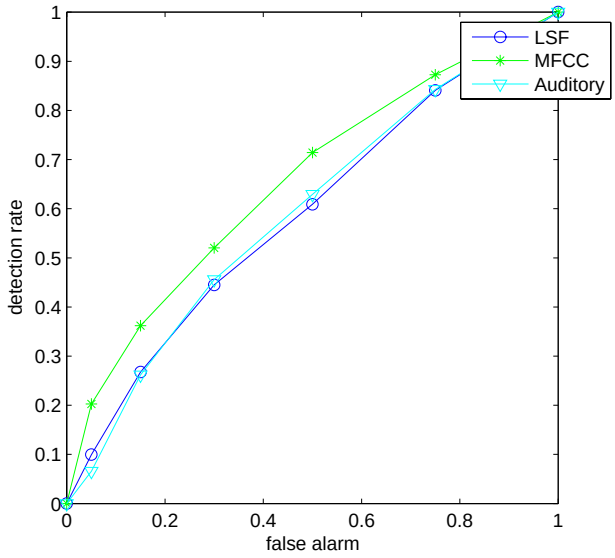
Figure 6.13: ROC curves for auditory model parameters, and various discriminant functions are shown.

and FM components was taken the detection rate for database No.1 was 43% for FLD and QD (Figure 6.12) which is the second highest detection rate for this database. For database No.2 detection rate was only 27% which is almost the half detection rate obtained from $a_k$ and $b_k$.

For the auditory case, the filterbank was chosen to have forty filters (or neurons). For each



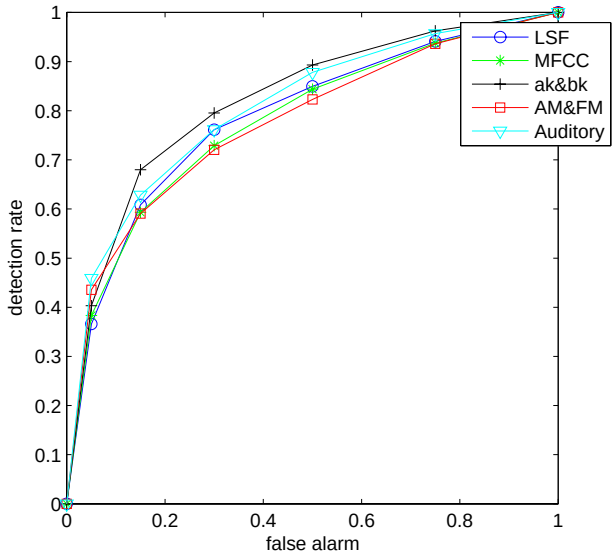

(a) Database No.1



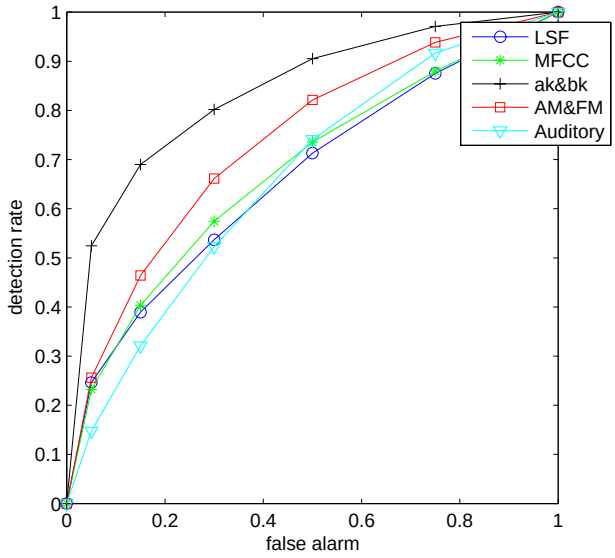

(b) Database No.2

Figure 6.14: ROC curves for Euclidean distance, and various features are shown.

neuron, mean neuron firing rate is taken from the cochleagrams as feature vector. Features from auditory model performed very well in database of Klabbers et al.. Actually, it gave detection rate of 50% (Figure 6.13), which was the highest detection rate for this database. In database of Stylianou et al., detection rate was much lower (only 20%), although, for false alarm 75%




(a) Database No.1

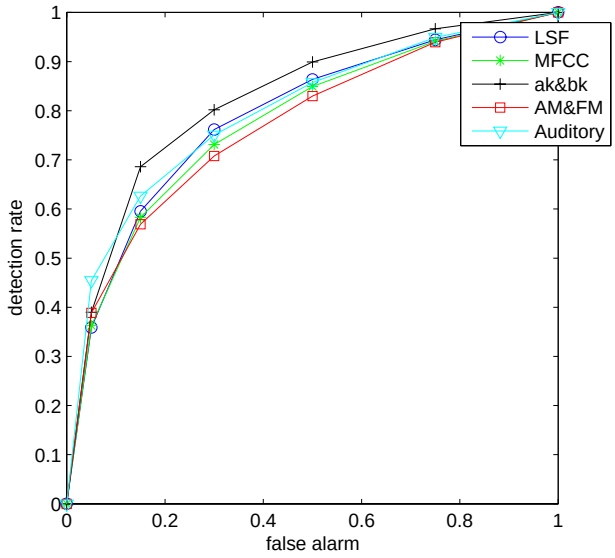



(b) Database No.2

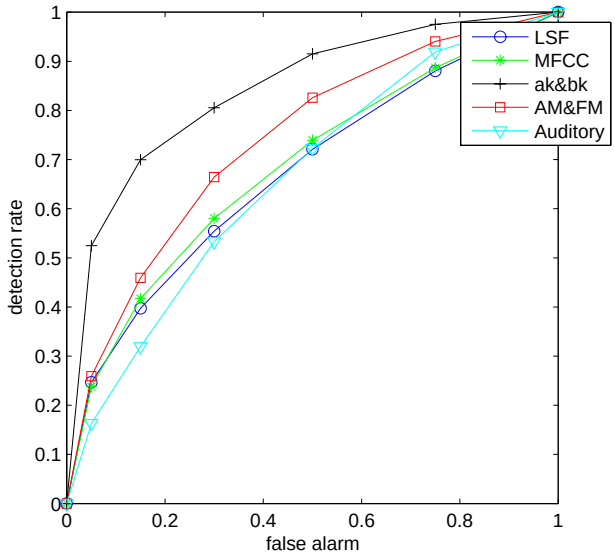
Figure 6.15: ROC curves for Fisher's linear discriminant, and various features are shown.

detection rate was 100% which means that there are not many cases where a discontinuous stimulus gave low cost.

Next, features are compared given a discriminant function. For $l_2$ norm we had "LSF < Auditory < MFCC" in database of Klabbers et al., while in database of Stylianou et al. we had




(a) Database No.1

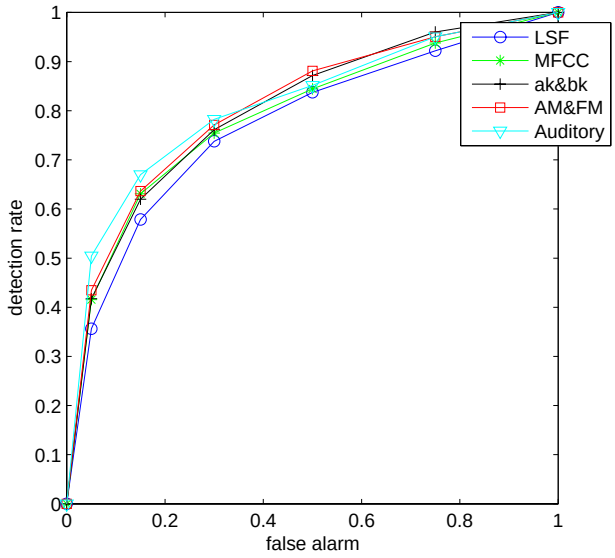



(b) Database No.2

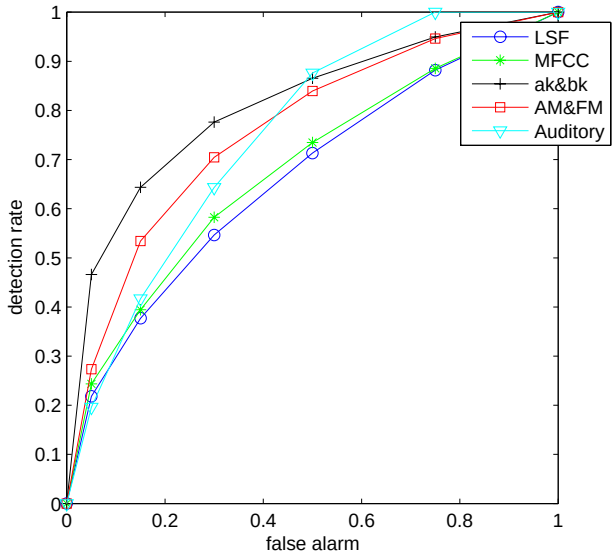
Figure 6.16: ROC curves for linear regression, and various features are shown.

"LSF = Auditory < MFCC" (Figure 6.14). Detection rates were rather low in both databases which means that a more sophisticated discrimination method is needed.

Using as discriminant method Fisher's linear discriminant detection rates are higher than $l_2$. For database No.2 nonlinear harmonic model performed extremely well compared to the other




(a) Database No.1




(b) Database No.2

Figure 6.17: ROC curves for quadratic discriminant, and various features are shown.

features (Figure 6.15). What is more, $a_k$ and $b_k$ provide high detection rates for database No.1 (detection rate was about 40% when false alarm was set to 5% and better to the other features for bigger false alarms). Exactly similar results can be obtained using linear regression method instead of Fisher's linear discriminant (Figure 6.16). This is expected since they have similar

properties.

Quadratic discriminant can be seen as a generalization of FLD. Therefore, it is natural to believe that it will give better results (Figure 6.17). However, this is not true because the need of estimating much more parameters makes our training samples being insufficient.

Tables that are following present the correlation coefficients in both databases for the various features and discrimination methods. In bold are the correlations that exceeds 0.5. Correlation coefficient values are in accordance with ROC curves as it is expected. The only exception is using QD. The basic reasons for this are that QD is not linear and not adequately trained.

| Distance | Database No.1 | Database No.2 |
|---|---|---|
| $l_1$ on LSF | 0.304 | 0.228 |
| $l_2$ on LSF | 0.295 | 0.180 |
| FLD on LSF | **0.535** | 0.318 |
| LR on LSF | **0.544** | 0.324 |
| QD on LSF | 0.365 | 0.234 |
| $l_1$ on MFCC | **0.506** | 0.297 |
| $l_2$ on MFCC | 0.495 | 0.315 |
| FLD on MFCC | **0.510** | 0.350 |
| LR on MFCC | **0.516** | 0.350 |
| QD on MFCC | 0.423 | 0.271 |

Table 6.1: Correlation coefficients for the LSF and the MFCC.

$a_k$&$b_k$ parameters using linear regression had the highest correlation coefficient in both databases. Then, $a_k$&$b_k$ parameters using FLD, auditory parameters using LR, AM–FM components using FLD or LR and LR on LSF follow for database No.1. In database No.2, there are not any high correlation coefficient except for $a_k$ of the nonlinear harmonic model.

| Distance | Database No.1 | Database No.2 |
|---|---|---|
| $l_1$ on $a_k$ | 0.110 | 0.321 |
| $l_2$ on $a_k$ | -0.272 | 0.556 |
| FLD on $a_k$ | **0.542** | **0.600** |
| LR on $a_k$ | **0.549** | **0.601** |
| QD on $a_k$ | 0.378 | 0.409 |
| $l_1$ on $b_k$ | -0.020 | -0.034 |
| $l_2$ on $b_k$ | -0.018 | -0.120 |
| FLD on $b_k$ | 0.434 | 0.229 |
| LR on $b_k$ | 0.431 | 0.224 |
| QD on $b_k$ | 0.302 | 0.184 |
| FLD on $a_k$&$b_k$ | **0.567** | **0.645** |
| LR on $a_k$&$b_k$ | **0.574** | **0.654** |
| QD on $a_k$&$b_k$ | 0.399 | 0.452 |

Table 6.2: Correlation coefficients for the nonlinear harmonic model.

| Distance | Database No.1 | Database No.2 |
|---|---|---|
| $l_1$ on AM | -0.026 | 0.252 |
| $l_2$ on AM | 0.121 | 0.252 |
| FLD on AM | 0.481 | 0.377 |
| LR on AM | 0.484 | 0.385 |
| QD on AM | 0.394 | 0.276 |
| $l_1$ on FM | -0.075 | 0.308 |
| $l_2$ on FM | -0.191 | 0.299 |
| FLD on FM | 0.405 | 0.364 |
| LR on FM | 0.415 | 0.371 |
| QD on FM | 0.426 | 0.328 |
| FLD on AM&FM | **0.522** | 0.450 |
| LR on AM&FM | **0.562** | 0.455 |
| QD on AM&FM | **0.515** | 0.395 |

Table 6.3: Correlation coefficients for the AM–FM components.

| Distance | Database No.1 | Database No.2 |
|---|---|---|
| $l_1$ on AuPa | 0.431 | 0.191 |
| $l_2$ on AuPa | 0.441 | 0.129 |
| FLD on AuPa | **0.549** | 0.304 |
| LR on AuPa | **0.562** | 0.307 |
| QD on AuPa | **0.515** | 0.394 |

Table 6.4: Correlation coefficients for the Lyon's auditory model. (AuPa = Auditory Parameters)

# Chapter 7

# Conclusions & Future Work

In the present work, we described several concatenation costs for the prediction/detection of discontinuities in concatenative speech synthesis systems. The evaluation of discontinuities was broken into two steps: *feature extraction* and *discrimination function*. This problem is really difficult because we do not know which acoustic cues the brain uses for the detection of a change in a speech signal.

Our feature sets come from speech analysis/synthesis, speech coding and speech recognition. Except for well known features such as LSF and MFCC, we presented nonlinear features from a *harmonic model with time varying amplitudes*, *AM-FM decomposition using DESA* and *Lyon's cochlea model*. In chapter 3, these five feature sets are analytically explained. The use of nonlinear features is supported from the nature of our ear system as well as the nature of the problem since concatenation of two signals is a nonlinear operation.

Distance measures or discrimination function are used for the comparison of two feature vectors and they are presented in Chapter 4. Metrics and statistical methods such as Fisher's linear discriminant and linear regression were used as discrimination functions. Statistical methods gave better results, however the small amount of data made training set small and for methods such as quadratic discriminant, which is a generalization of Fisher's linear discriminant, the output

results were not robust.

In Chapter 6, results obtained from two different psychoacoustic listening tests showed that nonlinear harmonic model using Fisher's linear discriminant or linear regression performed very well in both tests. It was significantly better than MFCC separated with Euclidean distance which a common concatenation cost in modern TTS systems. Another good concatenation cost, but less good than nonlinear harmonic model, is AM-FM decomposition again with Fisher's linear discriminant or linear regression. These results indicate that a concatenation cost which is based on nonlinear features separated by a statistical discriminant function is a good choice.

In the future, a collaboration with France Telecom is established and our methods will be tested to their concatenative speech synthesis system. Depending on the results, we will extend their system with our best concatenation costs.

# Bibliography


[1] David Klatt. Review of text-to-speech conversion for english. *Journal of the Acoustical Society of America*, 82:737–793, 1987.

[2] J. L. Flanagan. *Speech analysis synthesis and perception*. Springer, 1972.

[3] P. Taylor A. Black and R. Caley. The festival speech synthesis system. *www.cstr.ed.ac.uk/projects/festival.html*, 1998.

[4] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal. The AT&T Next-Gen TTS System. *137th meeting of the Acoustical Society of America*, 1999. http://www.research.att.com/projects/tts.

[5] Manfred R. Schroeder. A brief history of synthetic speech. *Speech Communication*, pages 231–237, 1993.

[6] M. S. Hunnicutt J. Allen and D. Klatt. *From text to speech*. Cambrigde University Press, 1987.

[7] Thierry Dutoit. *An Introduction to Text-to-Speech Synthesis*. Kluwer Academic Publishers, 1997.

[8] A. Hunt and A. Black. Unit selection in a concatenative speech synthesis system using large speech database. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 373–376, 1996.

[9] S. Isard and D. Miller. Diphone synthesis techniques. *Proceedings of IEE International Conference on Speech Input/Output*, pages 77–82, 1986.

[10] K. A. Lenzo and A. Black. Diphone collection and synthesis. *ICSLP 2000*, 2000.

[11] Robert E. Donovan. *Trainable Speech Synthesis*. PhD thesis, Cambridge University, Engineering Department, 1996.

[12] F. Violaro and O. Boeffard. A hydrid model for text-ot-speech synthesis. *IEEE Transactions on Speech and Audio Processing*, 6:426–434, Sep 1998.

[13] E. Klabbers and R. Veldhuis. Reducing audible spectral discontinuities. *IEEE Transactions on Speech and Audio Processing*, 9:39–51, Jan 2001.

[14] M. Tsuzaki T. Hirai, H. Kawai and N. Nishizawa. Analysis of major factors of naturalness degradation in concatenative synthesis. *Interspeech 2005*, pages 1925–1928, Sep 2005.

[15] N. Campbell and A. Black. *Procody and the selection of source units for concatenation synthesis*. Springer Verlag, 1995.

[16] J. Wouters and M. Macon. Perceptual evaluation of distance measures for concatenative speech synthesis. *International Conference on Spoken Language Processing ICSLP 98*, pages 2747–2750, 1998.

[17] E. Klabbers and R. Veldhuis. On the reduction of concatenation artefacts in diphone synthesis. *International Conference on Spoken Language Processing ICSLP 98*, pages 1983–1986, 1998.

[18] J.-D. Chen and N. Campbell. Objective distance measures for assessing concatenative speech synthesis. *EuroSpeech99*, pages 611–614, 1999.

[19] Y. Stylianou and A. Syrdal. Perceptual and objective detection of discontinuities in concatenative speech synthesis. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2001.

[20] Robert E. Donovan. A new distance measure for costing spectral discontinuities in concatenative speech synthesis. *The 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, 2001.

[21] J. Vepa S. King and P. Taylor. Objective distance measures for spectal discontinuities in concatenative speech synthesis. *ICSLP 2002*, pages 2605–2608, 2002.

[22] Jerome R. Bellegarda. A novel discontinuity metric for unit selection text-to-speech synthesis. *5th ISCA Speech Synthesis Worksop*, pages 133–138, 2004.

[23] Yannis Stylianou. *Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification*. PhD thesis, Ecole Nationale Supèrieure des Télécommunications, 1996.

[24] P. Maragos J. Kaiser and T. Quatieri. On separating amplitude from frequency modulations using energy operators. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Mar 1992.

[25] R. O. Duda P. E. Hart D. G. Stork. *Pattern Classification*. John Wiley & Sons, Inc, 2001.

[26] Y. Pantazis Y. Stylainou. Nonlinear speech features for the objective detection of discontinuities in concatenative speech synthesis. *Summer School in Nonlinear Speech Modeling and Applications*, pages 375–383, 2004.

[27] Y. Pantazis Y. Stylainou and E. Klabbers. Discontinuity detection in concatenated speech synthesis based on nonlinear analysis. *InterSpeech2005*, pages 2817–2820, 2005.

[28] Y. Pantazis Y. Stylainou. On the detection of discontinuities in concatenative speech synthesis. *Summer School in Nonlinear Speech Modeling and Applications*, 2005.

[29] J. Wouters and M. W. Macon. Unit fusion for concatenative speech synthesis. *ICSLP*, Oct 2000.

[30] J. H. L. Hansen and D. T. Chappel. An auditory-based distortion measure with application to concatenative speech synthesis. *IEEE Transactions on speech and audio processing*, 6:489–495, Sep 1998.

[31] W. N. Campbell and A. Black. Prosody and the selection of source units for concatenative synthesis. In R. Van Santen, R.Sproat, J.Hirschberg, and J.Olive, editors, *Progress in Speech Synthesis*, pages 279–292. Springer Verlag, 1996.

[32] R. Donovan and E. Eide. The ibm trainable speech synthesis system. *ICSLP*, 1998.

[33] J. Vepa and S. Taylor. Kalman-filter based join cost for unit selection speech synthesis. *Eurospeech*, Sep 2003.

[34] J. Vepa and S. King. Subjective evaluation of join cost and smoothing methods for unit selection speech synthesis. *IEEE Transactions on speech and audio processing*, 2005.

[35] O. Rosec C. Blouin, P. C. Bagshaw and C. d'Alessandro. Concatenation cost calculation and optimisation for unit selection in tts. *Proc. IEEE 2002 Workshop on Speech Synthesis*, Sep 2002.

[36] P. C. Bagshaw C. Blouin and O. Rosec. A method of unit pre-selection for speech synthesis based on acoustic clustering and decision trees. *ICASSP 2003*, pages 692–695, 2003.

[37] H. Kawai and M. Tsuzaki. Acoustic measures vs. phonetic features as predictors of audible discontinuity in concatenative synthesis. *ICSPL*, 2002.

[38] A. K. Syrdal and A. D. Conkie. Data-driven perceptually based join cost. *5th ISCA Speech Synthesis Workshop*, pages 49–54, 2004.

[39] A. K. Syrdal and A. D. Conkie. Perceptually-based data-driven join costs: Comparing join types. *Interspeech 2005*, pages 2813–2816, Sep 2005.

[40] F. K. Soong and B. H. Juang. Line spectrum pairs and speech data compression. *ICCASP*, pages 1.10.1–1.10.4, 1984.

[41] J. Markel and A. Gray. *Linear prediction of speech*. Springer Verlag, 1976.

[42] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoustics, Speech and Audio Processing*, 26:357–366, Aug. 1980.

[43] H. M. Teager. Some observations on oral air flow during phonation. *IEEE Trans. Acoust., Speech, Signal Processing*, Oct 1980.

[44] H. M. Teager and S. M. Teager. Evidence for nonlinear sound production mechanism in the vocal tract. *Speech Production and Speech Modelling*, 55, Jul 1990.

[45] James F. Kaiser. On a simple algorithm to calculate the 'energy' of a signal. *Proc. ICASSP-90*, pages 381–384, Apr 1990.

[46] P. Maragos T. J. Quatieri and J. F. Kaiser. Detecting nonlinearities in speech using an energy operator. *4th IEEE Digital Signal Precessing Workshop*, pages 19–20, Sep 1990.

[47] P. Maragos T. F. Quatieri and J. F. Kaiser. Speech nonlinearities, modulations and energy operators. *Proc. IEEE ICASSP-91*, May 1991.

[48] P. Cosi and E. Zovato. Lyon's auditory model inversion: a tool for sound separation and speech enhancement. *Proceedings of ESCA Workshop on 'The Auditory Basis of Speech Perceprion*, pages 194–197, Jul. 1996.

[49] R. F. Lyon. A computational model of filtering, detection, and compression in the cochlea. *Proceedings of IEEE-ICASSP-82*, pages 1282–1285, 1982.

[50] R. Patuzzi B.M. Johnstone and G.K. Yates. Basilar membrane measurements and the travelling waves. *Journal of Hearing Research*, 22, 1986.

[51] Malcolm Slaney. Lyon's cochlear model. *(Techn. Rep. # 1998-010)*, 1998.

[52] A. L. Edwards. *An Introduction to Linear Regression and Correlation*. W. H. Freeman, San Francisco, 1976.

[53] N.R. Draper and H. Smith. *Applied Regression Analysis*. Wiley Series in Probability and Statistics, 1998.

[54] A. S. House C. E. Williams M.H. L. Hecker and K. D. Kryter. Phycoacoustic speech test: A modified rhyme test. *Tech. Doc. Rept. ESD-TDR-63-403*, Jun 1963.

[55] Steven M. Kay. *Funtamentals of statistical signal processing: Detection theory.* Prentice Hall, 1998.