HELSINKI UNIVERSITY OF TECHNOLOGY
Department of Electrical and Communications Engineering
Laboratory of Acoustics and Audio Signal Processing

Petri Korhonen

Unsupervised Segmentation of Continuous Speech Using Vectorautoregressive Modeling

Master's Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Technology.

Espoo, December 13, 2004

Supervisor: Professor Unto K. Laine
Instructor: Professor Unto K. Laine

HELSINKI UNIVERSITY OF TECHNOLOGY

ABSTRACT OF THE MASTER'S THESIS

Author:	Petri Korhonen
Name of the thesis:	Unsupervised segmentation of continuous speech using vectorautoregressive modeling
Date:	Dec 13, 2004 **Number of pages:** 57
Department:	Electrical and Communications Engineering
Professorship:	S-89
Supervisor:	Prof. Unto K. Laine
Instructor:	Prof. Unto K. Laine

In this thesis a method for unsupervised segmentation of continuous speech is introduced. The method is based on Vector AutoRegressive (VAR) modeling. The VAR model is used in the auditory time-frequency domain to predict spectral changes. The forward and backward prediction error increases at the phone boundaries. These error signals are then used to study and detect the boundaries of largest changes allowing the most reliable automatic segmentation. The fully unsupervised method leads to segments consisting of a variable number of phones. The robustness and quality of performance of the method was tested by using a set of 201 Finnish sentences pronounced by one female and two male speakers. The boundaries between stops and vowels, in particular, are detected with high probability and accuracy.

Keywords: Speech analysis, speech segmentation, vector autoregressive modeling

TEKNILLINEN KORKEAKOULU DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Petri Korhonen	
Työn nimi:	Jatkuvan puheen automaattinen segmentoiminen käyttäen vektoriautoregressiomallinnusta	
Päivämäärä:	13.12.2004	**Sivuja:** 57
Osasto:	Sähkö- ja tietoliikennetekniikka	
Professuuri:	S-89	
Työn valvoja:	Prof. Unto K. Laine	
Työn ohjaaja:	Prof. Unto K. Laine	

Tässä diplomityössä esitellään menetelmä jatkuvan puheen segmentoimiseen. Menetelmä perustuu vektoriautoregressiiviseen (VAR) mallinnukseen. VAR mallia käytetään aika-taajuus alueen muutoksien ennustamiseen. Ennustus tehdään sekä mallia edeltävälle datalle, että mallin jälkeen tulevalle osalle. Mallin antama ennustusvirhe kasvaa äännerajoilla. Näitä virhesignaaleja käytetään segmenttirajojen havaitsemiseen. Suurimmat muutokset antavat luotettavimman segmentoinnin. Itseohjaava menetelmä tuottaa segmenttejä, jotka koostuvat vaihtelevasta määrästä äänteitä. Menetelmän häiriönsietokykyä ja laatua testattiin käyttäen 201 suomen kielen lausetta. Puhujina oli kaksi miestä ja yksi nainen. Erityisesti klusiilien ja vokaalien väliset rajat havaittiin luotettavasti ja tarkasti.

Avainsanat: Puheenanalyysi, vektoriautoregressio, puheen segmentointi

Acknowledgements

This Master's thesis has been a part of research project "Coding and Modeling of Phonemes in Speech" funded by Finnish Academy in the laboratory of Acoustic and Audio Signal Processing. The project was started under the TEKES funded USIX-STT project.

First and formost I would like to thank professor Unto K. Laine for his ideas, enthusiasm, and most of all the trust he has shown me over the years despite some difficult times in research. He has supported me in my work, but also he has helped me in my private life, making it possible for me to divide my time between Finland, Iceland and the U.S. in the past couple of years.

I would like to thank Jouni Pohjalainen, M.S.(Eng.) for letting me use his implementations of VAR model Matlab codes. Jouni has also helped me in numerous practical issues.

Working on a subject like this involves a lot of boring hours spent sitting in front of a computer screen. This thesis would not have been completed without my co-workers Carlo, Jykke, and Toni, who have made our office in the lab such a fun place to work in.

I also like to thank my parents for supporting me in my studies.

Finally, I would like to thank my love Catherine Kiwala for being there for me.

Otaniemi, December 13, 2004

Petri Korhonen

Contents

Abbreviations

AR	Autoregression
ASR	Automatic Speech Recognition
CVC	Conconant-Vowel-Conconant
DFT	Discrete Fourier Transform
DSP	Digital Signal Processing
MFCC	Mel Frequency Cepstral Coefficient
HMM	Hidden Markov Model
HSR	Human Speech Recognition
LP	Linear Prediction
LPC	Linear Predictive Coding
LSF	Line Spectrum Frequency
LSP	Line Spectrum Pair
STT	Speech-To-Text
TTS	Text-To-Speech
VAR	Vector Autoregression
WER	Word Error Rate
WLPC	Warped Linear Predictive Coding
WLSP	Warped Line Spectrum Frequencies

Chapter 1

Introduction

1.1 Motivation of this Thesis

Many subfields of speech technology need robust methods for automatic speech segmentation. Segmentations are indispensable for the initial training of acoustic ASR models, the development of TTS systems and speech research in general. Preferably, segmentation methods should be fully speaker- and language-independent. They should perform segmentation without any prior information about the speaker of the utterance in question. These methods should not rely on any type of prior learning, and should be able to process unknown utterances in a fully unsupervised manner.

In this thesis a novel method for automatic speech segmentation, which fulfills, to a certain degree, the hard demands mentioned, is introduced. The method is based on detecting unpredicted changes in auditory time-frequency representation of continuous speech at phone boundaries using Vector AutoRegressive modeling. The development of the segmentation system introduced here started from the aim of analyzing and capturing the temporal aspects of speech signal in more detail for speech recognition. The novel method presented in this thesis produces segments consisting of phone clusters of different lengths. The method does not find all the segment boundaries, since the some segment boundaries do not produce a rapid change in the spectra. When facing speaker independent unlimited vocabulary (e.g. inflectional languages) continuous speech recognition, the words have to be split into to smaller units such as morphemes; hence, not every phone boundary needs to be detected. Segments similar to syllables or morphemes consisting one to many phones do apply as well - as long as the total number of different segments is not too high for modeling purposes.

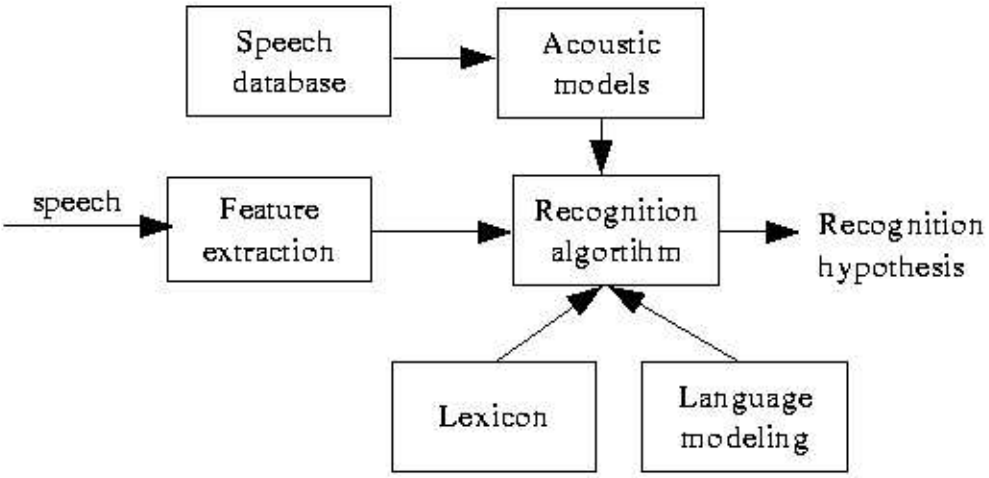

Figure 1.1: Block diagram of a speech recognition system

1.1.1 Segmentation for Speech Recognition Systems

Automatic recognition of unlimited vocabulary speaker independent continuous speech has been described as one of the most difficult engineering tasks at the moment. The main idea is simple: to transform spoken audio stream into a string of tokens i.e. text. Applications can vary from limited vocabulary speaker dependent word recognition to speaker independent unlimited vocabulary continuous speech recognition systems. The ultimate goal for speech recognition is to develop a system whose performance meets the speech recognition accuracy of humans. This task is by no means easy, and as the sophistication of the speech recognition systems grows, it becomes more obvious that we need methods specializing in subtasks. This thesis will decribe a solution for one of these subtasks in speech analysis.

The ways of performing speech recognition by machine can, broadly speaking, be divided into three classes: the acousic-phonetic approach, the pattern recognition approach, and the artificial intelligence approach [1]. Regardless of the methods used, most of the speech recognition systems share the same basic building blocks. The structure of state-of-the-art speech recognition is presented in Figure 1.1. The first block of a speech recognition system is a signal processing part (feature extraction). In frame-based speech recognition systems

the feature extraction part converts the time domain signal into a set of equally spaced discrete features usually computed at 10ms. In segment-based systems such as SUMMIT [2] the signal is broken into variable length segments corresponding the hypothesized phones. Each segment has a feature vector representing the phone. Using the segmental framework for recognition, the richer set of acoustic-phonetic features can be used. In both frame-based and segment-based systems these features are supposed to carry compact yet sufficient information about speech for classification into units. These units can be, for example, phonemes, diphones, triphones, words, or even longer units. The choice of the recognition unit depends on the application, and even on the language, which is the case in the Finnish language, where the number of possible words poses limitations for practical purposes. The feature extraction part is followed by the recognizer. The actual recognizer part uses acoustic models, lexicon, and language modeling to produce a recognition hypothesis from these features [3].

1.1.2 ASR and HSR

Despite the significant advances made in the field of speech research, especially since the advent of HMM based ASR systems, this goal is nowhere near completion. The comparison of speech recognition made by machines and humans was conducted 1997 by Lippmann [4]. Since there have only been incremental improvements to HMM based ASR systems, this comparison can still be considered valid [5]. Lippmann's conclusion is that to reduce the gap between ASR and HSR, the most effort should go into to improvement of the low level acoustic-phonetic modeling. His studies showed that the performance of the ASR system on a continuous speech corpus drops from 3.6% WER to 17% WER when grammar information is not used. The corresponding reduction of WER with HSR was from 0.1% to 2%. ASR performance on a connected alphabet task was reported to be about 4% while the same error rate with HSR was 1.6%. This 1.6% WER can be considered as an upper bound of human performance on an isolated alphabet. This shows that the ASR is much more dependent on high level language information than HSR. Most improvements in ASR systems in recent years have notably been in the field of language modeling.

Perceptual experiments carried out by Fletcher [5][6] give more evidence that humans carry out highly accurate phoneme level recognition. The recognition error of phones in nonsense consonant-vowel-consonant (CVC) syllables in best conditions was reported to be 1.5%. It was also reported that the probability of correct recognition of the syllable is a product of the recognition of the constituent phones. In review of Fletcher's work, Allen [7] inferred that individual phones must be correctly recognized for a syllable to be recognized correctly. His conclusion was that it is unlikely that context is used in the early stages of

speech recognition. This suggests that the focus in ASR research must be on phone recognition. Fletcher also suggests that the recognition is done in separate frequency bands, with recognition error rate being the minimum of error rates across all the frequency bands. In HMM-based systems, the recognition is always carried out using all the frequency bands at the same time. This makes the HMM based systems very different from human speech recognition. Furthermore, state-of-the-art HMM based ASR systems use triphones as the basic recognition unit, because of the poor performance of HMMs on the phoneme level. Also, HMM based systems do not take into account the temporal information, which is a distinctive feature in Finnish.

1.1.3 Shortcomings of Frame Based HMM Systems

HMMs have been the most widely used paradigm for speech recognition. They are higly effective themselves, but some properties of HMMs can be questioned in use for speech processing; especially for phonetic speech recognition [8]. Firstly, in most HMM-based ASR systems, the acoustic modeling is restricted to an observation space defined by a temporal sequence of feature vectors computed at a fixed frame rate. Within the same phonetic segment, the adjacent frames often exhibit smooth dynamics and frames that are highly correlated. This violates the conditional independence assumption by the HMM. However, the relationship between frames computed in different phonetic segments is weaker. This could motivate a framework that makes fewer conditional independence assumptions between observation frames within a phonetic segment. This implies that we need a way to extract the phonetic segments from a speech signal. Another property of the HMM-based ASR systems is that they use homogenous, frame-based feature vectors such as Mel-frequency cepstral coefficients (MFCCs) to represent speech. This might not be enough to capture certain acoustic measurements known to be important for phonetic distinctions. Acoustic cues that best characterize the phonetic distictions are tied to temporal landmarks in a speech signal. These are, for example, the point of oral closure or release, or other points of closure or opening in the vocal tract produced during speech production. Laine and Hirvonen [9] showed that temporal information is important with recognition of stop consonants. They state that the stop consonants, which are clearly the most difficult sounds for a speech recognizer, can be recognized with close to 100% accuracy with proper design of classifier. This requires knowledge of the timing information of the sounds. These temporal landmarks often correspond to phonetic boundaries, which has made many speech reseachers consider segment- and landmark-based approaches for automatic speech recognition.

1.2 Outline of the thesis

This thesis is divided into six chapters. Chapter 1 is laying the foundation for the thesis. In chapter 2, the human speech production and perception chain are shortly reviewed. Also in chapter 2 we take a look at the nature of the speech signal itself. In chapter three the signal processing methods for speech analysis that are used in this work are introduced. Also in chapter three, we shortly take a look at the methods others have proposed for automatic speech segmentation. In chapter four the proposed segmentation algorithm is introduced. In chapter five we evaluate the performance of the proposed segmentation method, comparing it to segmentation conducted by human transcribers. Chapter six is devoted to conclusions and perspectives, and it also lays down path to the future improvements of the proposed method.

Chapter 2

Speech as a Tool for Communication

2.1 Introduction

In order to adequately model natural speech, we need to utilize knowledge about spoken language structure. Essentially, the chain of a spoken language system can be divided into production and perception, which are equally important components in this chain. The chain can be further divided into distinct elements, as depicted in Figure 2.1.

Spoken language is fundamentally used to communicate information from a speaker to a listener. It begins with a thought and intent to communicate in the brain. This semantic

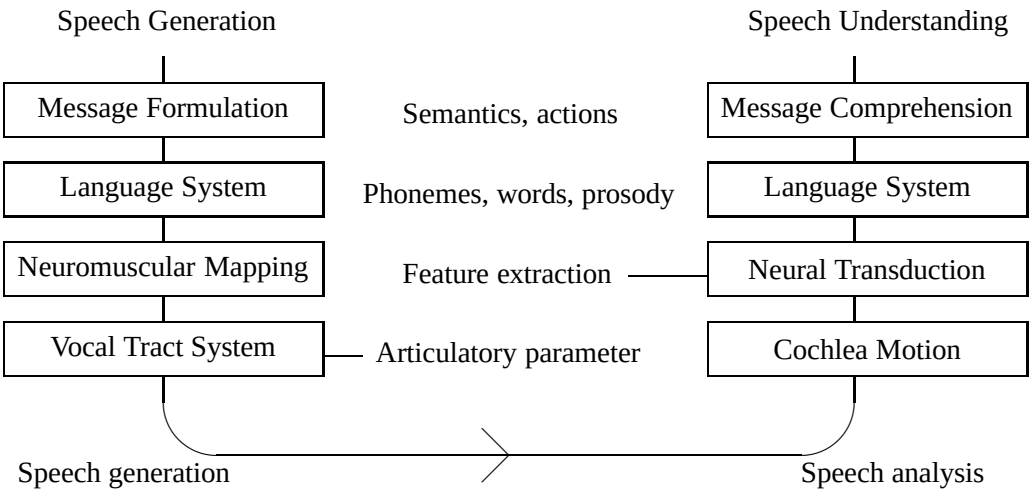

Figure 2.1: Chain of speech communication from production to perception

message in a person's mind should be transmitted to the listener via speech. The next step is to convert this message into a sequence of words. Each word is considered to consist of a sequence of phonemes that corresponds to the pronunciation of words. The term *phonemes* is used to denote any of the minimal units of speech sound in a language that can serve to distinguish one word from another.

Speech is not merely a sequence of phonemes; each sentence also contains a prosodic pattern that denotes the duration of each phoneme, intonation of the sentence, and the loudness of sounds. After the language system has carried out this mapping, the talker executes a series of neuromuscular signals. These neuromuscular commands perform articulatory mapping to control the vocal cords, lips, jaw, tongue, and velum. These vocal organs together produce the sound sequence as the final output. [10]

In the perception part of the spoken language system, the speech understanding works in reverse order. First the sound is passed through the outer- and middle-ear to the cochlea in the inner ear. Inner ear performs frequency analysis as a filter bank. This spectral signal is converted into activity signals on the auditory nerve by neural transduction process. From that point on it is currently unclear how neural activity is mapped into the language system and how message comprehension is achieved in the brain. [10]

2.2 Speech Production and Acoustic-Phonetics

2.2.1 Anatomy and Physiology of Speech Organs

Speech sounds are produced using speech organs. Here their structure and function are shortly reviewed.

Organs of speech production can be divided into three parts: *subglottal system*, *larynx and its surroundings* and *supraglottal system*. These are presented in Figure 2.2. In this chain, airflow from the lungs passes through the larynx and vocal tract. This airflow exits the mouth as pressure variations constituting the speech signal. [11]

The Subglottal System

The subglottal system consists of the lungs and the thorax. The lungs are situated in the chest or thorax cavity. They are the source of an airflow that flows through the larynx and vocal tract. In all speech sounds, the basic source of power is the respiratory system, which pushes air out of the lungs.

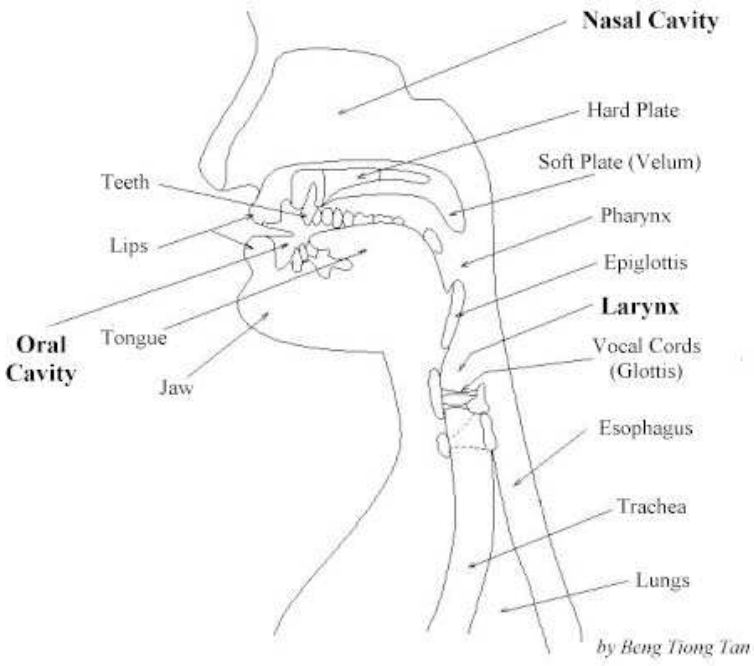

Figure 2.2: Vocal organs

The Larynx and Vocal Folds

The larynx constitutes of four cartilages: *thyroid*, *cricoid*, *arytenoid* and *epiglottis*. These cartilages are joined by ligaments and membranes. The larynx connects the lungs to the vocal tract through a passage called the *trachea*. Within the larynx there are a pair of elastic structures of tendon, muscles, and mucous membrane called vocal folds. The length, thickness and position of vocal folds can be controlled by means of various muscle contractions. During normal breathing, the vocal folds remain sufficiently parted to allow free air passage with little audible sound. The source of voiced speech sounds occurs in the larynx where vocal folds obstruct airflow from lungs partially or completely to create either turbulent noise or pulses of air. [12]

The Supraglottal System

The vocal tract is the most important component in the speech production system, since it provides the means to produce the many different sounds that characterize spoken language. It has two speech functions. Firstly it can shape the spectral distribution of energy in the glottal sound wave. Secondly, it can contribute to the generation of *obstruent* sounds. [12] The vocal tract can be modeled as an acoustics tube with resonances, called formants, which can be altered by moving articulators.

2.2.2 Articulatory phonetics

Articulatory phonetics relates linguistic features of sounds to positions and movements of the speech organs. Though humans can produce an infinite number of sounds[1], each language has a set of abstract linguistic units, called phonemes, to describe its sounds. A *phoneme* is defined as the smallest contrastive unit in the phonology of a language [12](p. 55). The sounds within each phoneme class usually have some articulatory gesture(s) or configuration(s) in common. Each language typically has 20-40 phonemes. In the Finnish phoneme set there are eight vowels and thirteen consonants. Also three additional consonants appear in loanwords.[3]

Consonants can be grouped a in such way that the same pattern or feature forms a description class for the group. Each class consists of consonants which are produced the same way. These classes are called *articulatory features*. There are three types of articulatory features of consonants: (1) features of manner of articulation (2) the voicing feature and (3) features of place of articulation.

Manner of Articulation

Manner of articulation is concerned with the way air flows through the vocal tract, which path(s) it takes, and the degree it is impeded by vocal tract constrictions.

Vowels are produced when air flows through, the oral cavity meeting no constriction in the middle of the vocal tract narrow enough to cause turbulent flow. In finnish there are 8 vowels (/a, e, i, o, u, y, ae, oe/). All the phonemes that do not match the criteria of vowels are consonants. Compared to consonants, the production mechanism of different vowels is quite similar. There are large differences between consonant classes, and thus conconants are here divided into six subclasses: stops, fricatives, nasals, tremulants, laterals, and semivowels.

Stops are produced with a complete or almost complete closure of a vocal tract and subsequent release of obstruction. After a vocal tract closure, the pressure builds up behind the occlusion. At this point there is little or no sound present. The sudden release of this pressure creates a brief (e.g., 10 ms) acoustic *burst* or *explosion*. Stops can be either voiced or unvoiced. Voicing is maintained during the closure of voiced stops. The voicing requires air to enter the vocal tract behind the occlusion, expanding the tract until the occlusion is

[1] within the constraints of the vocal tract

released. During this expansion sound is weakly radiated through the walls of the vocal tract. Voiced and unvoiced pairs of stop consonants include /b,p/, /d,t/, and /g,k/.

Nasals are always voiced consonants which are produced by letting the air flow through the nasal cavity by lowering the velum, and closing the vocal tract.

Fricatives are produced through constriction somewhere in the vocal tract, in the pharynx (rarely) or at the glottis narrow enough to produce noisy turbulent air flow. In the Finnish language the fricatives are voiceless, except /h/ which can be voiced between two vowels. There are two fricatives in the Finnish language [h,s] and phoneme [f] can be found in loan words.

Laterals. The vocal tract is blocked by pressing the tip of a tongue against the alveolar ridge. However, there is a passage on both sides of the tip of a tongue to let sound waves and air flow freely. The finnish language has one lateral /l/.

Semi-vowels are much like vowels, but the constriction of the vocal tract is more powerful, less stable, and more context dependent. This are /j,v/

Trill. In the Finnish language /r/ is produced by letting the tip of the tongue vibrate against the alveolar ridge. The rate at which the tongue vibrates is typically 20 to 25 Hz. The vibration produces an effect similar to amplitude modulation.

Place of Articulation

In order to produce consonants, the airstream from the lungs must be obstructed. Consonants can be classified according to the place and manner of these obstructions. Place of articulation is concerned with the point of narrowest vocal tract constriction that enables finer discrimination of phonemes. Phonemes according to their places of articulation are called bilabials, labiodentals, dentals, alveolars, retroflex, palato-alveolars, palatals, and velars [13]. Places of articulation are shown in Figure 2.3

2.2.3 Grapheme to Phoneme Conversion

Graphemes are the set of written symbols that are used to represent speech. Graphemes include for example letters, Chinese ideograms, numerals, punctuation marks, and other symbols. In writing systems that use letters as unit, the graphemes are grouped together in a string to represent words. They are not completely arbitrary, but have some correspondance

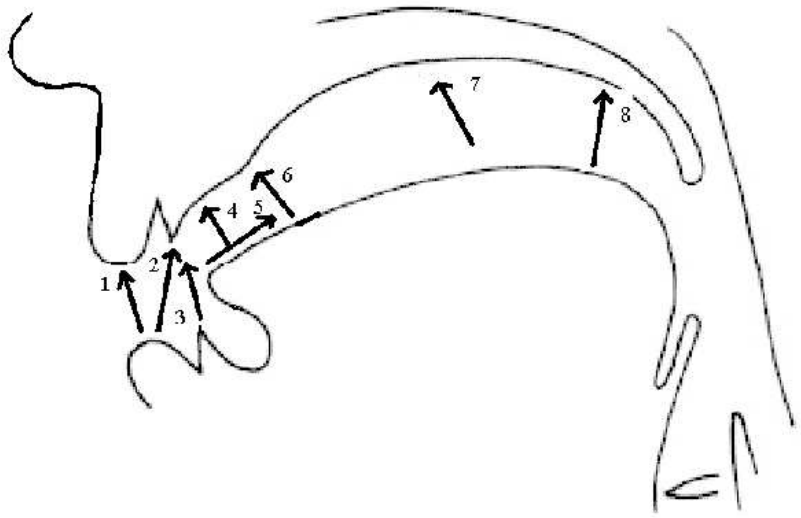

Figure 2.3: Places of articulation. 1. Bilabial; 2. Labiodental; 3. Dental; 4. Alveolar; 5. Retroflex; 6. Palato-Alveolar; 7. Palatal; 8. Velar.

to the phonemes in words. Unlike most languages of the world, the Finnish language's grapheme to phoneme conversion is quite straightforward. There is one to one mapping between graphemes and phonemes, with only a few exceptions.

2.3 Hearing

Hearing function is performed by the organ we call the ear, but recent research has emphasized that much of our hearing also depends on the data processing that occurs in the central nervous system as well [14]. This knowledge of nonlinearities of human hearing is more and more taken into account in speech processing.

2.3.1 Anatomy and Physiology of the Ear

The anatomy of ear is often divided into three sections: the outer ear, the middle ear and the inner ear. Here anatomy and functions of these three parts are shortly reviewed.

Outer ear. The parts of the outer ear are the pinna and auditory canal (meatus) which is terminated by the eardrum (tympanum). The function of the outer ear is to contribute to determination the direction of origin of sounds. Also, the auditory canal is a 1/4-wave pipe resonator, which boosts frequencies from 2000Hz to 5000Hz. The pinna funnels sound waves into the ear canal, and then sounds are conducted to middle ear via eardrum.

Middle ear. The middle ear consists of three small bones called ossicles (hammer, anvil, stirrup), which are connected to outer ear via the eardrum. The ossicles act as a lever, which changes pressure exerted by a sound wave on the eardrum into greater pressure on the oval window of the inner ear. Also the difference in the areas of eardrum and oval window boosts the signal.

Inner ear. The inner ear is composed of semicircular canals and the cochlea. The semicircular canals are the body's horizontal-vertical detectors, but contribute little or nothing to hearing. The cochlea contains all the mechanisms for transforming pressure variations into properly coded neural impulses. It is connected to the middle ear through the oval and round windows.

2.3.2 Sound Perception

Hearing works in a non-linear fashion. The physiological behavior of the ear in response to simple tones is relatively straightfoward. However, most sounds are time varying and have many spectral components. The cochlear processes of basilar membrane vibration and neural firings are highly nonlinear. As a result the perception of sound energy at one frequency is dependent on the distribution of sound at other frequencies and on the time course of energy before and after the sound. What a person hears in response to a given sound is an often complicated question.

Nonlinear Frequency Scales

Psychoacoustic experimental work has been undertaken to derive frequency scales that attempt to model the natural response of the human perceptual system. This is due to the fact that the cochlea of the inner ear acts as a spectrum analyzer. Perceptual attributes of sounds at different frequencies may not be entirely simple and linear in nature. Pitch is a term used to describe the position of sound on a scale from high to low. It is a subjective sensation. The pitch of a sound is mainly determined by the frequency for pure tones, but pitch may also change with sound level. For complex sounds, it also depends on the spectrum of the sound and duration.

The Mel-scale is a perceptually motivated frequency scale based on experiments with sinusoids. In these experiments subjects were required to divide given frequency ranges into four perceptually equal intervals or, alternatively, to adjust the frequency of a stimullus tone to be half as high as that of a comparison tone. The Mel-scale is linear below 1 kHz, and logarithmic above that. The Mel-scale is hoped to model sensitivity of human ear more

closely than a purely linear scale and to provide for greater discriminatory capability between speech segments. Mel-scale frequency analysis has been widely used in automatic speech recognition systems. Equation 2.1 can be used to approximate the Mel-scale.

$$B(f) = 1125 \ln(1 + f/700) \tag{2.1}$$

Another perceptually motivated scale is the Bark frequency scale. The Bark-scale is based on critical bands. The auditory system performs frequency analysis of sounds into component frequencies. The cochlea in the inner ear acts as overlapping filters having bandwidths equal to the critical bandwidth. As with the Mel-scale, it is also hoped that treating spectral energy over the Bark scale, a more natural fit with spectral processing in the ear can be achieved. Bark frequency can be calculated from linear frequency (in Hz) with equation 2.2 [15].

$$b(f) = 13 \arctan(0.00076f) + 3.5 * \arctan((f/7500)^2) \tag{2.2}$$

From these two scales Mel scale has gained popularity in the ASR community.

2.4 Speech Perception

2.4.1 Introduction

Much is known about how audition converts speech signals into patterns of auditory nerve firings, but the mechanisms by which the brain translates these nerve firings into a linguistic message are much less understood. Underneath the apparent ease to understand speech easily under most conditions lurk complex processes. One way to appreciate this complexity is to consider attempts to develop computer systems for automatic speech recognition. As impressive as the developments these systems have undergone are, they pale in comparison to the ability of humans to understand the speech in different settings.

2.4.2 Problems Posed By The Speech Stimulus

The researchers who are trying to understand the mechanisms of speech perception face the problem that the relationships between the acoustic signal and the sounds we hear are extremely complex. There are two main reasons for this complexity: namely, the segmentation problem and the variability of the acoustic signal. These problems are the reasons why it has been so difficult to design machines that can recognize spontaneous continuous speech.

The Segmentation Problem

If we listen to someone speak, we can segment the stream of speech we hear into individual words quite easily. If we take the same acoustic signal, we can see that the signal is not neatly separated into individual words or phonemes. The acoustic signal is continuous and there are not necessarily any clear physical breaks in the signal, or there might be breaks that do not correspond to the breaks we perceive between words or phonemes.

Sometimes the context is needed to achieve the correct segmentation, because two different words can have the same phonetic structure. For example ”I scream”, and ”ice cream” are identical, and the different word segmentation must be achieved by the meaning of the sentence in which the words appear.

Phonetic segmentation of speech signal is essentially a task of determining a time instance where one phoneme changes to another. This task is relatively easy if in the speech production there is some discrete even that marks the begining of a pronunciation of a phoneme. For example the time instance of a release burst is clearly defined event in the temporal domain. On the other hand if we consider for example a transition from one vowel to another, there is no discrete physical event that happens in speech production, but instead the vocal organs move rather slowly from one target position to another. To determine where one phoneme ended and another started is in this case is rather difficult.

The Variability Problem

The phoneme can have different forms that are determined by variety of sources. This phenomenon has been called “acoustic-phonetic non-invariance problem” by Klatt [16]. If we consider phonemes to have a perceptual reality, they might be expected to possess some acoustic characteristics that serve to differentiate one phoneme from another. Convincing evidence of existence of such phonetic invariance has not been found for all phonemes despite careful research.

Variability from a Phoneme’s Context Depending on its context, properties of the acoustic signal associated with a phoneme changes. The effect of context is a result of the way speech is produced. As we speak the articulators are constantly moving, so the shape of the vocal tract for a particular phoneme is influenced by the shapes for the preceding and following phonemes. This phenomenon (overlap between the articulation of neighboring phonemes) is called coarticulation.

Variability from Different Speakers For different speakers a particular phoneme can have very different acoustic signals. People say the same phonemes and words with different pitch, different accents, and different speed. Also the fine structure of vocal organs change from one person to another. Speakers also introduce variability to acoustic signal with sloppy pronunciation.

The variability in the acoustic signal caused by reasons explained in the previous paragraphs creates problems for the listener. Highly variable speech signals should be transformed into familiar words. Because of the segmentation problem and the variability problem, it has been difficult to design machines that can recognize speech. On the other hand humans somehow have the ability recognize speech despite the effects which seem to cause extreme difficulties for machines. It is to some extent unclear how human do it, but research conducted over the past 50 years has begun to unravel the mystery surrounding the human speech perception. It is not to say we need to build a machine that performs the exact same functions as human perception, but it might help us in the creation of an machine capable to understand spoken language.

2.4.3 Vowel Perception

The production of steady-state vowels can be described in terms of static vocal tract shapes. These shapes provide prototype targets for vowels, when they are articulated in words. The perception of vowels in the isolated case (i.e no coarticulation effects from neighboring phones) is based on their steady-state spectra. The locations of first three formants (F1-F3) are considered to be most important factors[2].

In case of diphthongs, the shape of the spectra in the endpoint steady states is important. The formant frequency locations for each vowel depends on three factors: the length of the pharyngeal-oral tract, the location of constrictions in the tracts, and the degree of narrowness of the constrictions.

In reality, the process of vowel perception is a much more complex process. We perceive different vowel patterns of men, women, and children as the same vowel, and constant vowels are perceived despite changes in its formant pattern due to coarticulation in different phonetic contexts. The changing formants, over their trajectories between consonants,

[2]Formant is a resonance of the vocal tract. It is common to use the word formant to describe the spectrum peaks, but strictly speaking the formants are acoustical properties of vocal tract that produced the spectrum and only the effects of the formants are seen in the spectrum pattern of a speech, because spectrum is strongly affected by the resonances of the vocal tract.

are the key information that is processed into a constant vowel perception. Acoustic signal associated with diphthongs could have steady-state formant pattern, if the speaker would sustain vowels by maintaining the articulatory target in a "steady-state" for a longer period of time. However, in fluent speech, the vocal tract is typically in constant motion, and thus formant patterns also change throughout the course of syllables. This suggest that natural the unit for automatic recognition of continuous speech could consist of a sequence of several phonemes (e.g. syllable). The segmentation in syllabic level would obviously be easier than in the phonetic level.

Perception of Steady-State Vowels

Acoustical analysis of natural vowels have shown that center frequencies for formants F1-F3 vary systematically for different vowels. Perceptual studies have, however, shown that vowels can be synthetically simulated by two-formant patterns. In these two-formant patterns, one or the other formant constitutes a weighted average of two or more formants. It seems that when formant frequencies are close, as F1 and F2 are for back-vowels, and F2 and F3 are for front vowels, they are perceptually integrated to form the "effective" formant, which equivalent to an average of the two close spectral peaks. This integration happens when two spectral peaks occur within a critical distance in a psychophysical frequency scale (Bark scale). This effective formant is a single spectral peak of weighted average, both in frequency and amplitude, of the cluster of the formants within a range of 3.0-3.5 Barks. When center frequencies of formants exceed 3.5 Barks they are perceptually distinct. This in turn means that for drawing a distinction between vowels in automatic systems, the model for speech spectrum does not have to be highly detailed, especially if auditory scales are used.

Because formant frequencies are determined by the size and shape of vocal tract, the absolute values for same vowels produced by different speakers are not the same. Most obvious is the variation between men, women, and children. These differences can be so big, that not only we have great deal of variation of F1 and F2 within a vowel category, but there can be overlap between different vowel categories (i.e sometimes different vowels have same F1 and F2 frequencies). Still we can tell what vowel the speaker intended to produce. This phenomenon of perceiving the same vowel for different acoustical realizations is called "speaker normalization problem".

Perception of Coarticulated Vowels

It can easily be demonstrated that steady-state acoustic targets for vowels are not often reached in spontaneous speech. In consonant vowel context vowel formant frequencies are shifted from target values produced in null contexts. These shifts cause “shrinking” of F1/F2 space, such that formant patterns of different vowels are more similar to each other. This means a reduction in the acoustic contrast among vowels when they are produced in consonantal context. The place of articulation of preceding and following consonant results in different amounts of reduction.

These acoustic effects of coarticulation are referred to as *target undershoot*. Speakers’ articulatory intentions are the same for all contexts, but as these articulatory intentions are carried out at increasingly rapid rates, the speech organs fail to to reach the positions that they assume when the vowel is pronounced under ideal steady-state conditions, because of the physiological limitations of speech organs. With respect to the perception of coarticulated vowels, listeners compensate for target undershoot in order to to recover canonical vowel targets. Listeners show “perceptual overshoot” in identification of vowels in CVC syllables with moving formant patterns. Perception of coarticulated vowels is not based exclusively on the frequencies of first three formants of the vowel in the nucleus of syllable where they approach their canonical values most closely, but also the direction and slope of formant transitions into and out of the syllable nucleus affect the perceived identity of the vowel [17]. This would also support the syllable as a recognition unit as opposed to phoneme.

2.4.4 Consonant Perception

As it was shown in Chapter 2.2 consonants are produced by rapid articulatory gestures that are superimposed on the slower, more global movements for the vowels. Consonants form syllable units with vowels, where vowels are the syllable nuclei and consonants occur at the onsets and offsets of syllables. Consonant gestures make temporary constrictions in the vocal tract. The vocal tract can be narrowed to cause turbulent, noise-like sound, or even to block it completely. Consonant gestures can be interpreted as marking the syllable borders. Because the consonants are produced by reconfiguration of the vocal tract shape between consonants and vowels, the sound patterns associated with consonants involve changes in the formants. Also abrupt bursts of noise and/or silences are often involved with consonants.

2.5 Speech as Patterns on Paper

In the previous sections we showed that the spectral structure of speech signal plays a pivotal role in speech perception. Distinctions between different phonemes is based often on the difference in speech spectrum. Is it possible for a computer to find phoneme segment boundaries from a stream of fluent speech using spectral information? To give us some idea if it can be done, and how well, we can look at how well humans perform this task. Ever since the invention of the sound spectrograph nearly 60 years ago, there have been numerous attempts to study spectrogram reading. The spectrogram often does not provide adequate information on certain lingustically relevant cues, such as stress and intonation. Nevertheless, it gives a good description of the segmental acoustic cues of speech.

In 1980, Cole et. al. examined in detail the methods used by a highly skilled expert spectrogram reader [18]. Their qualitative and quantitative studies showed that in the task of phoneme level segmentation, the expert identified 97% of the segments defined by transcribers. Their study also focused on the methods that the expert was using for placing the segment boundaries. The expert appeared to make use of only a few simple priciples, as shown in the following protocol excerpt:

> *"I am marking at various places*
> *where it shows, you know,*
> *maximal spectral difference...*
> *I'm basically using the spectral change*
> *as a parameter for marking the boundaries...*
> *There is an intensity,*
> *a sharp intensity difference..."*

Spectral changes accompany changes in manner of articulation, which in turn are characteristic for each phone. A succession of phones will produce successive changes in the speech spectrum, and the expert places segment boundaries at these points of change.

It has been argued that, on theoretical grounds, spectrogram reading cannot be learned, because the speech signal is such a complex code that phonemes can only be perceived through the working of a special decoder. The results the human were able to achieve are nevertheless encouraging, and suggests that it is possible for a machine to achieve at least 97% accuracy for segmentation, just by investigating the spectral changes of a speech signal; the acoustic signal is the primary information-bearer.

It is interesting to note that the method that the human used to visually segment the speech

signal was based on the acoustic patterns that are apparently speaker independent and are not degraded by the use of a spectrographic representation. A spectrogram-like representation of a speech signal thus would appear to be adequate for high quality speech systems.

Chapter 3

Preprocessing of Speech Signals

3.1 Autoregressive models for speech analysis

3.1.1 Linear prediction

A very powerful method for speech analysis is based on linear prediction (LP). In the field of speech research the method is called LP-modeling, but in the other fields is called autoregressive modeling (AR). It is a fast, simple and effective way of estimating the main parameters of speech signals. It gives a precise representation of the speech spectral magnitude. The drawback of the LP is that, to minimize analysis complexity, the speech signal is usually assumed to come from an all-pole source. This means that model spectrum has no zeros, though the actual speech spectrum has zeros due to the glottal source as well as zeros from the vocal tract response in nasals and unvoiced sounds.

Linear prediction provides an analysis-synthesis system for speech signals. The synthesis model consists of an excitation source $U(z)$ that gives input to a spectral shaping filter $H(z)$. Output of this system is speech $\hat{S}(z)$. $U(z)$ and $H(z)$ are chosen, so that $\hat{S}(z)$ is as close as possible in some sense to the original speech $S(z)$. Usually $U(z)$ is chosen to have a flat spectral envelope so that spectral detail is confined to $H(z)$. This choice is a reasonable assumption since the excitation for unvoiced sounds resembles white noise. The source for voiced sounds is viewed as a uniform sample train, having a line spectrum with uniform-area harmonics. In reality vocal cord puffs of air can be be modeled as the output of a glottal filter whose input is the sample train. The spectral shaping effects of both vocal tract and the glottis are combined into one filter $H(z)$. [12][10][19]

$H(z)$ is obtained for speech signal $s(n)$ by first windowing signal for frames of N samples. In this window the signal is concidered to be stationary. This allows the filter $H(z)$ to be

modeled with constant coefficients. In general *pole-zero* case, $H(z)$ is assumed to have p poles and q zeros. In this case speech sample $\hat{s}(n)$ can be modeled by a linear combination of the p previous output samples and $q+1$ previous input samples

$$\hat{s}(n) = \sum_{k=1}^{p} a_k \hat{s}(n-k) + G \sum_{l=0}^{q} b_l (n-l) \tag{3.1}$$

where G is a gain factor for the input speech (assuming $b_0 = 1$). We can specify the above in the frequency domain by taking the z transform on both sides of equation 3.1.

$$H(z) = G \frac{1 + \sum_{l=0}^{q} b_l z^{-l}}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{3.2}$$

In practice most LP work assumes all-pole model ($q = 0$). Thus we get

$$H(z) = G \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{3.3}$$

where G is a gain factor and a_i are LP-model coefficients. If speech $s(n)$ is filtered by an inverse filter (predictor filter) $A(Z)$ (the inverse of an all-pole $H(z)$

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} \tag{3.4}$$

the output $e(n)$ is called *error* or *residual* signal:

$$e(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \tag{3.5}$$

Now the $E(z)$ and $U(z)$ should be similar, so that $H(z)$ would model the vocal tract system response. Since speech production cannot be fully modeled with an all-pole filter, there is differences between $e(n)$ and $u(n)$.

Obtaining filter coefficients for LP-model The set of LP coefficients a_k characterizing an all-pole $H(z)$ model of the speech spectrum can be obtained using the classical least-squares method. This method chooses the filter coefficients a_k to minimize the mean energy in the error signal over a frame of speech data. In the autocorrelation method (windowing method) of the least-squares technique, the speech signal is multiplied by a Hamming or similar time window to obtain signal with finite duration

$$x(n) = w(n)s(n) \tag{3.6}$$

The length of the window $w(n)$ is typically choosen to be 20 to 30 ms during which the signal $s(n)$ is assumed to be stationary. Windowing makes $x(n) = 0$ outside the range

$n = 0 \leq N - 1$.

For residual $e(n)$ corresponding the windowed signal $x(n)$ the energy E is

$$E = \sum_{n=-\infty}^{\infty} e^2(n) = \sum_{n=-\infty}^{\infty} [x(n) - \sum_{k=1}^{p} a_k x(n-k)]^2 \tag{3.7}$$

Now the task is to choose values a_k that minimize residual energy by setting $\partial E/\partial a_k = 0$ for $k = 1, 2, \ldots, p$. This leads to p linear equations in p unknowns a_k.

$$\sum_{n=-\infty}^{\infty} x(n-i)x(n) = \sum_{k=1}^{p} a_k \sum_{n=-\infty}^{\infty} x(n-i)x(n-k), \quad i = 1, 2, \ldots, p \tag{3.8}$$

Noticing that the first term of equation 3.8 is autocorrelation $r(i)$ of $x(n)$ and taking advantage of the finite duration of $x(n)$ leads to

$$\sum_{k=1}^{p} a_k r(i-k) = r(i), \quad i = 1, 2, \ldots, p \tag{3.9}$$

where autocorrelation terms are $r(i) = \sum_{n=-\infty}^{n=\infty} s_n s_{n-1}$. The autocorrelation function is even: $r(-i) = r(i)$. Also since s_n is nonzero only during the N samples, is is sufficient to compute only

$$r(i) = \sum_{n=1}^{N-1} s_n s_{n-i}, \quad 0 \leq i \leq p \tag{3.10}$$

The Equation 3.9 can be written in matrix form as

$$\begin{pmatrix} r(0) & r(1) & r(2) & \ldots & r(p-1) \\ r(1) & r(0) & r(1) & \ldots & r(p-2) \\ r(2) & r(1) & r(0) & \ldots & r(p-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r(p-1) & r(p-2) & r(p-3) & \ldots & r(0) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} r(1) \\ r(2) \\ r(3) \\ \vdots \\ r(p) \end{pmatrix} \tag{3.11}$$

or equally with corresponding symbols

$$\boldsymbol{Ra} = \boldsymbol{r} \tag{3.12}$$

The LP coefficients can be computed with

$$\boldsymbol{a} = \boldsymbol{R^{-1}r} \tag{3.13}$$

It is worth noting that the autocorrelation matrix R is symmetric and Toeplitz matrix, the coefficients can be computed without inverting the autocorrelation matrix using Levinson-Durbin recursion. The gain can be computed with

$$G = \sqrt{R(0) - \sum_{k=1}^{p} a_k r(k)} \tag{3.14}$$

3.1.2 Different representations for LP

LP analysis produces a vector of *p* real-valued coefficients a_k. These coefficients represent an optimal estimate to the spectrum of the windowed speech using p poles. This same information, can be represented in many different formats. These other formats may be more useful or more physically interpretable than others. These formats include for example reflection coefficients, impulse response $h(n)$ of the LP synthesis filter $H(z)$, autocorrelation coefficients for a_k or $h(n)$, spectral coefficients from a DFT of either autocorrelation, the cepstrum of a_k or $h(n)$, the log-area ratios, inverse sine functions and line spectrum pair (LSP).

The line spectrum pairs are used in the segmentation algorithm introduced in this thesis, and thus in the following sections they are discussed in more detail.

LSP

The representation for the LP parameters that will be examined here in more detail is called *line spectrum pair* (LSP). The method was first introduced by Itakura [20], and it has since become a widely utilized and investigated method for representing LP parameters. The procedure for converting LP-coefficients to LSPs involves mapping the *p* zeros of $A(z)$ onto the unit circle through two z-transforms $P(z)$ and $Q(z)$ of $(p+1)$st order:

$$P(z) = A(z) + z^{-(p+1)}A(z^{-1}) \tag{3.15}$$

$$Q(z) = A(z) - z^{-(p+1)}A(z^{-1}) \tag{3.16}$$

And directly from equations 3.15 and 3.16 it follows that

$$A(z) = \frac{[P(z) + Q(z)]}{2} \tag{3.17}$$

The roots of the polynomials $P(z)$ and $Q(z)$ are called the LSPs.

LSP's have some interesting properties. Firstly, the zeros of $P(z)$ and $Q(z)$ lie on the unit circle. Secondly, it can be shown that zeros of $P(z)$ and $Q(z)$ alternate as ω increases along the circle. Also the LSP coefficients correspond to the frequencies of these zeros. The LSP coefficients allow interpretation in terms of formant frequencies because each complex zero of $A(z)$ maps into one zero in each of $P(z)$ and $Q(z)$. If these zeros have close frequencies, it is likely that the original $A(z)$ corresponds to a formant; otherwise, the zero is likely to be of wide bandwidth and to contribute only to the tilt of the spectrum [12]. This last property gives the motivation for LSPs' usage for the segmentation method described later in this thesis.

3.1.3 Frequency Warping

As it was shown in Chapter 2.3 human auditory system works in highly non-linear fashion. It is also time-variant, and adaptive in many ways. Combined, these properties make models of auditory perception complex, and audio techniques that utilize these priciples are intricate. There are only a few properties of the auditory system that can be exploited in audio signal processing easily and systematically. Probably the most common auditory feature that is taken into account in audio signal processing are the pitch scales explaned in Chapter 2.3.2.

Frequency warping is a process of transforming one spectral representation for signals on some frequency scale to another representaion on a new frequency scale. The original scale (e.g., Hz, f-domain) has a certain frequency resolution; most often uniform, and the new frequency scale is non-uniform (e.g., Bark). The new representation obtained through this transform has a uniform frequency resolution on a new scale. However, when observed from the original frequency scale, it has a nonuniform frequency resolution. This transformation from one frequency scale to another is presented with the warping function $v(f)$ which defines the relation between the new frequency scale and the old one. The warping function $v(f)$ is a smooth and monotonic function.

Warped linear prediction (WLPC) dates back to 1980 (Strube) [21], but the idea of performing linear predictive analysis on a modified frequency scale have been introduced earlier (for example see [22]). The frequency scale used in digital signal-processing has conventionally been linear in relation to hertz (Hz) scale, i.e., frequency resolution is uniform for the whole band from dc to Nyquist frequency ($0.5f_s$, where f_s is sampling frequency). The basic building block of DSP is unit delay z^{-1}, which delays signal components of all frequencies by the same amount. There exist ways to implement DSP on a warped freqency scale that approximates Bark scale instead of uniform frequency scale [21]. One technique to implement frequency warped DSP algorithms is to replace unit-delays with first-order all-pass filters to obtain a variable digital filter that can be controlled by adjusting the coefficient of the all-pass element. Transfer function of first-order all-pass filter is given by

$$D(z) = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}} \tag{3.18}$$

By definition, an all-pass filter has a constant magnitude response for the whole frequency band. However, the phase response of $D(z)$ varies with λ. With $\lambda = 0$, the transfer fuction reduces to a single unit delay linear phase and constant group delay. Now we can build an all-pass filter chain by cascading a set of all-pass filters. If λ is positive, the

nonuniform group delay makes low-frequency components of a signal proceed slower and high-frequency components faster than in a chain of unit delays. The mapping between natural frequency domain, and the warped frequency domain is determined by the phase function of the all-pass filter, given by

$$\omega' = \arctan \frac{(1-\lambda^2)\sin(\omega)}{(1+\lambda^2)\cos(\omega) - 2\lambda} \tag{3.19}$$

where $\omega = 2\pi f/f_s$, and f_s is the sampling frequency. For a specific value of λ, it is possible to achieve the frequency transformation that closely resembles the frequency mapping occuring in the human auditory system. [23] It would be desirable to design signal processing systems and algorithms that work directly on some auditory frequency scale. In this chapter we review a general approach to designing DSP techniques on a warped frequency scale that approximates the Bark scale as presented in [23].

Warped Linear Prediction

In the previous chapter the classical forward linear prediction was shown. In classical LP analysis an estimate for the sample value $x(n)$ was obtained as a linear combination of N previous values given by

$$\hat{x}(n) = \sum_{k=1}^{N} a_k x(n-k) \tag{3.20}$$

and same written in z-plane

$$\hat{X}(z) = (\sum_{k=1}^{N} a_k z^{-k}) X(z) \tag{3.21}$$

where a_k $k = 1, \ldots, N$ are fixed filter coefficients. Here the shift operator z^k (unit delay filter or a shift operator) may be replaced by a first-order all-pass filter. Denoting all-pass filter by $D(z)$ we can write the equation 3.1.3 as

$$\hat{X}(z) = [\sum_{k=1}^{N} a_k D(z)^k] X(z) \tag{3.22}$$

$D(z)^{-k}$ in equation 3.1.3 can be interpreted as a generalized shift operator defined as

$$d_k[x(n)] \equiv \delta(n) \star \delta(n) \star \cdots \star \delta(n) \star x(n) \tag{3.23}$$

where the asterisk indicates convolution and $\delta(n)$ is impulse response of the filter $D(z)$. $\delta(n)$'s form a k-fold convolution. In $k = 0$ let us write $d_0[x(n)] \equiv x(n)$. The error of the estimate to be minimized may now be written as

$$e = E\{|x(n) - \sum_{k=1}^{N} a_k d_k[x(n)]|^2\} \tag{3.24}$$

where $E\{\cdot\}$ denotes expectation. Filter coefficients a_k are solved minimizing this error. Setting $\partial e/\partial a_k$ with $k = 1, 2, \ldots, N$ leads to system of normal equations,

$$E\{d_m[x(n)]d_0[x(n)]\} - \sum_{k=1}^{N} a_k E\{d_k[x(n)]d_m[x(n)]\} = 0 \tag{3.25}$$

where $m = 0, \ldots, N-1$. It is shown that since $D(z)$ is an all-pass filter, the equation [23]

$$E\{d_m[x(n)]d_k[x(n)]\} = E\{d_{m+p}[x(n)]d_{k+p}[x(n)]\} \tag{3.26}$$

holds, for all values of m, k, and p. This means that in both parts of equation 3.1.3 there appear the same correlation values. This in turn means that equation 3.1.3 can be seen as a generalized form of the Wiener-Hopf equations. Figure 3.1 shows an autocorrelation network structure, which can be used to solve the correlation terms. The optimal filter coefficiants a_k can be solved efficiently using the Levinson-Durbin algorithm, like in the case of classical autocorrelation method of linear prediction.

Now we have a prediction error filter given by

$$A(z) = 1 - \sum_{k=1}^{N} a_k D(z)^k \tag{3.27}$$

This filter can be implemented easily by simply replacing all unit delays (Z^{-1}) of a conventional FIR structure with all-pass filter $D(z)$ blocks. The LP synthesis filter (IIR)

$$A^{-1}(z) = \frac{1}{1 - \sum_{k=1}^{N} a_k D(z)^k} \tag{3.28}$$

is possible to implement using techniques discussed in [23][24]. [23] Using frequency warped methods for linear prediction the order of the model can be significantly lower compared to the conventional LP-modeling. This is of great importance when used with vector autoregression, which will be expained in more detail in next chapter.

3.2 Vector Autoregression

Finite order vector autoregressive (VAR) models are used in forecasting and structural analysis. There has not been extensive use of these models in the field of speech processing. Some attempts to model temporal variations of speech spectra with vector autoregression has been done, for example, in the field of speaker identification. [25] [26] [27] [28]

Here first the models are introduced and later the least squares estimation procedure to solve the estimation problem is introduced.

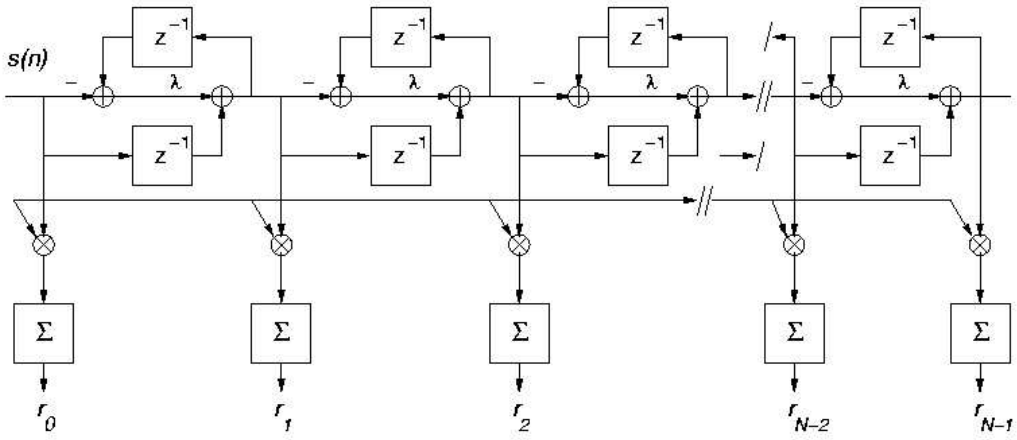

Figure 3.1: Warped autocorrelation network

3.2.1 Stable VAR(p) Processes

Assume a K-dimensional multiple time series $\mathbf{y}_1, \ldots, \mathbf{y}_T$ where $\mathbf{y}_t = (\mathbf{y}_{1t}, \ldots, \mathbf{y}_{Kt})'$ and that the series is known to be generated by a stationary and stable VAR(p) process

$$\mathbf{y}_t = \mathbf{v} + \mathbf{A}_1\mathbf{y}_{t-1} + \ldots + \mathbf{A}_p\mathbf{y}_{t-p} + \mathbf{u}_t. \tag{3.29}$$

In equation 3.29 $\mathbf{y}_t = (y_{1t}, \ldots, y_{Kt})'$ is a $(K \times 1)$ random vector, the $\mathbf{A}_i$ are fixed $(K \times K)$ coefficient matrices, $\mathbf{v} = (v_1, \ldots, v_K)'$ is a fixed $(K \times 1)$ vector of intercept terms allowing for the possibility of nonzero mean $E(\mathbf{y}_t)$ and $\mathbf{u}_t = (u_{1t}, \ldots, u_{Kt})'$ is a K-dimensional *white noise* with non-singular covariance matrix $\mathbf{C}_u$. The coefficients $\mathbf{v}, \mathbf{A}_1, \ldots, \mathbf{A}_p$, and $\mathbf{C}_u$ are unknown parameters. The coefficients will be estimated from the multivariate time series data. The VAR(1) is called stable if all eigenvalues of $\mathbf{A}_1$ have modulus less than 1.

Multivariate Least Sequares Estimation

Assume the time series $\mathbf{y}_1, \ldots, \mathbf{y}_T$ of $\mathbf{y}$ variables to be available. Also we assume p pre-sample values for each variable, $\mathbf{y}_{-p+1}, \ldots, \mathbf{y}_0$ to be available. To simplify the notation let

us define

$$\mathbf{Y} := (\mathbf{y}_1, \ldots, \mathbf{y}_T) \qquad (K \times T),$$

$$\mathbf{B} := (\mathbf{v}, \mathbf{A}_1, \ldots, \mathbf{A}_p) \quad (K \times (Kp+1)),$$

$$\mathbf{Z}_t := \begin{bmatrix} 1 \\ \mathbf{y}_t \\ \vdots \\ \mathbf{y}_{t-p+1} \end{bmatrix} \qquad ((Kp+1) \times 1),$$

$$\mathbf{Z} := (\mathbf{Z}_0, \ldots, \mathbf{Z}_{T-1}) \quad ((Kp+1) \times T),$$

$$\mathbf{U} := (\mathbf{u}_1, \ldots, \mathbf{u}_T) \qquad (K \times T),$$

With the notation defined above we can write the VAR(p) model for $t = 1, \ldots, T$ as

$$\mathbf{Y} = \mathbf{B}\mathbf{Z} + \mathbf{U} \tag{3.30}$$

It has been shown in [29] that in least squares sense the estimation for $\mathbf{B}$ can be obtained as

$$\hat{\mathbf{B}} = \mathbf{Y}\mathbf{Z}'(\mathbf{Z}\mathbf{Z}')^{-1} \tag{3.31}$$

In this thesis models of order $p = 1$ are used only (i.e., VAR(1)-models).

3.3 Approaches to Automatic Speech Segmentation

In this section will give a short review of the existing methods for automatic speech segmentation. The segmentation systems can be broadly divided into two categories. One class of algorithms perform the segmentation when both the signal and the underlying sequence of phonemes is known [30]. Also some of these systems require either manually or automatically segmented training data. Another class of algorithms take only the speech signal as an input, and do not use any knowledge about the underlying phoneme sequence contained within the speech signal [31].. Instead these algorithms locate the time instances where segment boundaries are, based on where there is a high degree of variation in speech waveform. The system proposed in this thesis is unsupervised segmentation method that takes as input only the speech signal, and does not require training data.

The output of the segmentation systems can vary from simple end-point detection, to phoneme

or syllable level segmentation depends on the application in question.

The segmentation systems can also be divided into two categories based on the approach they use for segmentation. Some systems are based on specific acoustic cues or features for the segmentation [5][32][33]. They typically focus, for instance, on transient behavior or specific differences between phoneme classes. The other category of segmentation systems use general features and acoustic modeling which are common in ASR systems [34]. It is not in the scope of this thesis to discuss all the different types of segmentation algorithms, so here we will just mention some attempts to perform segmentation using acoustic cues without the use of phonetic transcription.

In 1975 Mermelstein [35] proposed a convex hull algorithm to segment speech into syllabic length units. The algorithm used maxima and minima in a loudness measure extracted from the speech signal to find the prominent peaks and dips. The peaks were marked as syllabic peaks and the points near the syllabic peaks with maximal difference in loudness were marked as syllable boundaries.

The approach adopted by Sharma et. al. [36] suggested a procedure involving finding the ”optimal” number of sub-word segments in the given speech sample, before locating the sub-word segment boundaries. They used dynamic programming and subjective loudness function to achieve this. As with many segmentation systems developed in the past, the results shown in this paper are not comparable with other results, because there is not any absolute ”right” solution for segmentation.

Aversano et. al. [37] proposed a method for segmentation using *jump-function*. *Jump-function* is a measure of variation of a local spectrum. It is formed by comparing the averages of spectrum on both sides of hypothesised segment boundaries. A peak detection algorithm was introduced to select the peak values of *jump-function* to represent the hypothesized segment boundaries.

In a recent paper Prashad et. al. (2004) [38] present a method to automatically segment speech into syllable length units. Their method is based on “group delay processing” of the short time energy function of the continuous speech signal. At segment boundary they have achieved an error rate less than 20% of the syllable duration for 70% of the syllables. Their system introduces 5% insertions.

Chapter 4

Algorithm

4.1 Segmentation Algorithm

4.1.1 Introduction

In this chapter the proposed segmentation algorithm will be described in detail. The algorithm is based on analyzing variations in speech spectrum with vector autoregressive modeling (VAR). The speech signal is batch processed in non-real time. The method does not use labeled data to teach the system in any way; the only input for the algorithm is the utterance to be segmented. The systems based on this algorithm are speaker independent (some parameters might need adaptation for individual speakers). The aim of the algorithm is to efficiently, and accurately, find the time instances from the speech signal, where there is a change in short-time the speech spectrum. It is expected that the biggest and most rapid changes in the speech spectrum have resulted from transitions from one segment to another segment. Studies so far have shown that this segmentation corresponds, to some degree, to manual segmentation done by trained phoneticians, based on phonetic boundaries. The first order vector autoregressive model described in the chapter 3.2 tries to predict the vector at time t, using the vector at time $t-1$. The model is estimated from a sequence of vectors with the least squares estimate (LSE) method. The error used for estimation of the model is the one-step prediction error between subsequent vectors. Essentially, the model is used to produce values of the vector at time t from the values of vector at time $t-1$. In this thesis first order VAR models are used exlucively (VAR(1)).

$$\hat{\mathbf{y}}_t = \mathbf{A}\mathbf{y}_{t-1} + \mathbf{v} \tag{4.1}$$

Where $\mathbf{y}_t$ is a $(K \times 1)$ vector, $\mathbf{A}$ is fixed $(K \times K)$ coefficient matrix, and $\mathbf{v}$ is a fixed $(K \times 1)$ vector allowing for the possibility of nonzero mean $E(\mathbf{y}_t)$. This property of the model can be further exploited, by using the model to *recursively* produce vectors for values

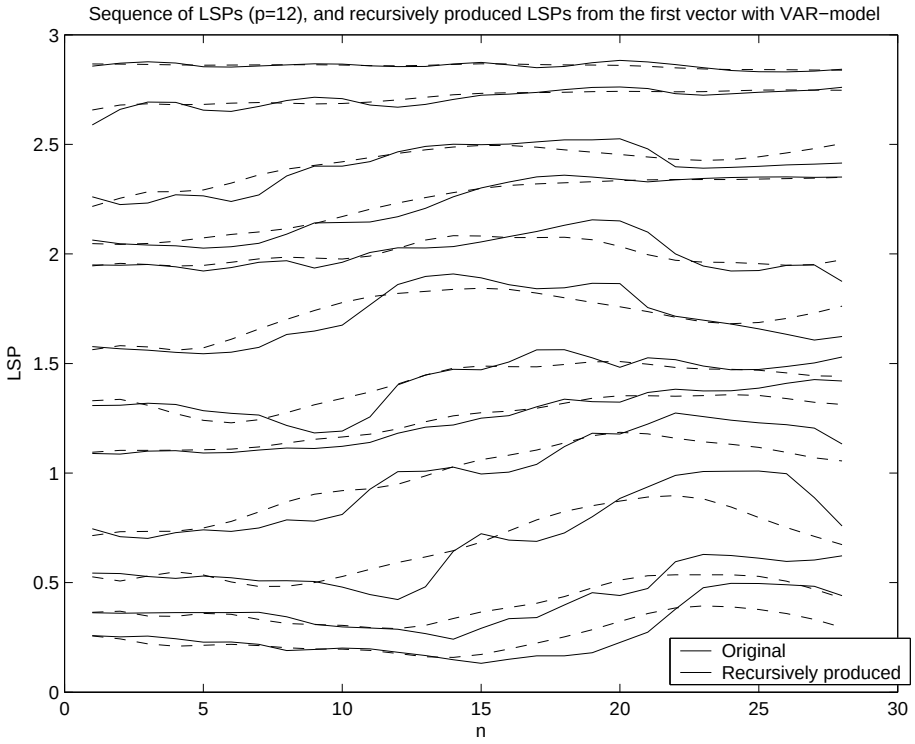

Figure 4.1: Recursively produced multivariate time series

$t+1, t+2 \ldots t+M$.

$$\hat{\mathbf{y}}_{t+1} = \mathbf{A}\hat{\mathbf{y}}_t + \mathbf{v} \tag{4.2}$$

In Figure 4.1, an example of recursively produced vector sequence is shown. The original multivariate time series is shown with solid lines, and the recursively produced times series with dashed line. First order VAR model was estimated from the original time series using least squares estimation. Then a new time series was created using the first vector of the original sequence with equation 4.2. From the figure we can see that the simple first order VAR model follows the trends of the original time series, but does not follow the fine structure of the original signal. If we estimate a model from vectors taken from a steady part of an utterance (e.g. long vowel) preceding a segment boundary, we can expect that model to produce similar vectors with the vectors in the steady part, when used to recursively produce new vectors. It also means that the difference e_t between the original vectors $\mathbf{y}_t$, and the recursively produced vectors $\hat{\mathbf{y}}_t$, over the boundary, should be large. The prediction error can be defined as

$$e_t = \sum^{p} (\mathbf{y}_t - \hat{\mathbf{y}}_t)^2 \tag{4.3}$$

where p is the length of the vector $\mathbf{y}_t$. This model sees only the past, and cannot predict the changes the original signal will have in the future. The model keeps producing steady part vectors, regardless of the changes in the original signal. This error between the original vectors and the recursively-produced vectors is used to detect spectral changes at segment boundaries in the algorithm presented here.

4.1.2 The Algorithm

First the digital speech signal $s(n)$ is converted into a sequence of short-time features $\mathbf{y}_t$ with each being a $(p \times 1)$ vector, where t is the frame that the coefficients were computed from. In this work, the short time features used are p:th order frequency warped line spectrum pairs (WLSP) ($p = 12 - 16$). The motivation for using WLSP's can be found from the chapter 3.1.2. The short-time features should be computed with relatively short intervals to obtain adequate time resolution for the purpose. This requires a lot of overlap between the subsequent frames. The step size used in this work is 3ms as opposed to 10ms which is the standard in most ASR systems. Also this step-size maches better to the time resolution of human hearing. Let $\mathbf{A}_t$ be the VAR(1) model computed from the data preceding the frame at time t

$$\mathbf{A}_t = VAR_{LSE}(\mathbf{y}_{t-L+1}, \ldots, \mathbf{y}_t) \tag{4.4}$$

The value of L should correspond to average length of a steady state of a phoneme in speech. For each vector $\mathbf{y}_t$ we compute recursively M estimates with the models $\mathbf{A}_{t-M} \ldots \mathbf{A}_{t-1}$

$$\begin{aligned} \hat{\mathbf{y}}_{t1} &= \mathbf{A}_{t-1}\mathbf{y}_{t-1} \\ \hat{\mathbf{y}}_{t2} &= \mathbf{A}_{t-2}^2\mathbf{y}_{t-2} \\ &\vdots \\ \hat{\mathbf{y}}_{tM} &= \mathbf{A}_{t-M}^M\mathbf{y}_{t-M} \end{aligned} \tag{4.5}$$

From these estimates we can compute relative errors

$$\begin{aligned} e_{t1} &= \frac{(\mathbf{y}_t - \hat{\mathbf{y}}_{t1})^T(\mathbf{y}_t - \hat{\mathbf{y}}_{t1})}{\mathbf{v}_t^T \cdot \mathbf{v}_t} \\ e_{t2} &= \frac{(\mathbf{y}_t - \hat{\mathbf{y}}_{t2})^T(\mathbf{y}_t - \hat{\mathbf{y}}_{t2})}{\mathbf{v}_t^T \cdot \mathbf{v}_t} \\ &\vdots \\ e_{tM} &= \frac{(\mathbf{y}_t - \hat{\mathbf{y}}_{tM})^T(\mathbf{y}_t - \hat{\mathbf{y}}_{tM})}{\mathbf{v}_t^T \cdot \mathbf{v}_t} \end{aligned} \tag{4.6}$$

From these errors we select the median value to represent the error at time t

$$e_t = \mathsf{median}(e_{t1}, \ldots, e_{tM}) \tag{4.7}$$

The small values of e_t are emphasized taking the logarithm of the error signal

$$E_t = 10 * \log_{10}(1 + e_t) \tag{4.8}$$

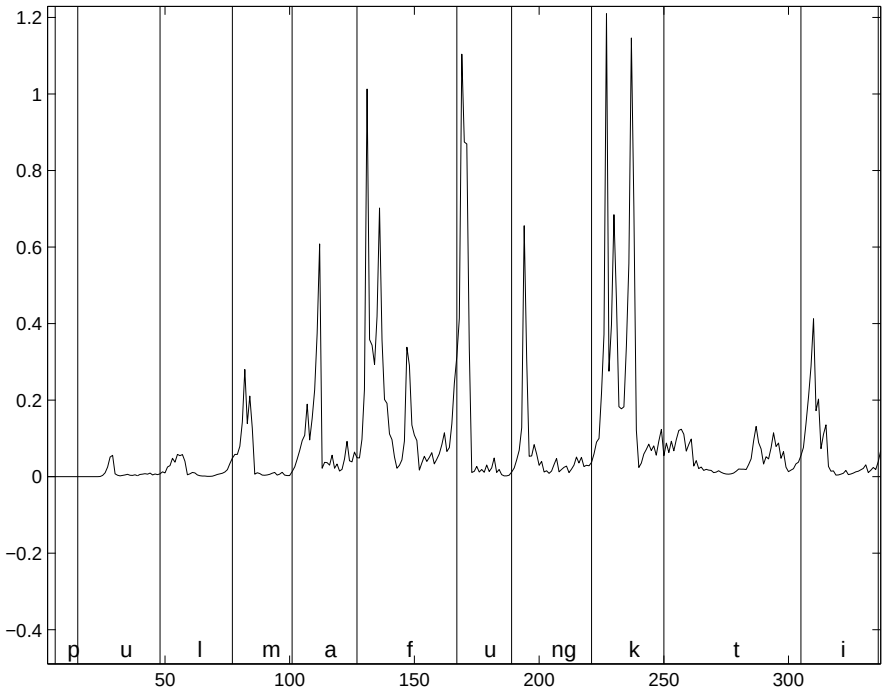

Figure 4.2: An error signal E_t over one sentence. The vertical lines are manually assigned segment boundaries

E_t is a signal that has large values, whenever there has been a steady state, followed by a sudden change in the speech spectrum. An example of error signal during one sentence is shown in Figure 4.2. The vertical lines in Figure 4.2 are boundaries of phonemes set by a trained phonetician. Model $\mathbf{A}$ is used to predict the values for $\mathbf{y}_t$ outside the window, from which the model was estimated. Until this point the model $\mathbf{A}_t$ has been used to recursively produce vectors for time instances $t \ldots t + M$. That means, that model predicts the *future* values of $\mathbf{y}_t$. The model can be used to predict the values outside of its scope also for values *before* the model. This can be easily done by reversing the original sequence of vectors $\mathbf{y}$, and performing the same analysis as we have done so far for the reversed signal. After doing the analysis for the reversed signal also, we have two error signals, which we will denote by E_{t+} and E_{t-}: forward and backward prediction error. These two signals are presented in Figure 4.3. To help the visualization the backwards error has been negated. In the next step of the algorithm, the errors E_{t+} and E_{t-} are combined to a single error E_{t*} with

$$E_{t*} = E_{t+} - E_{t-} \tag{4.9}$$

The resultant error should have a large negative peak before a segment boundary, and large positive peak after a segment boundary. An example of combined error signal E_{t*} is shown

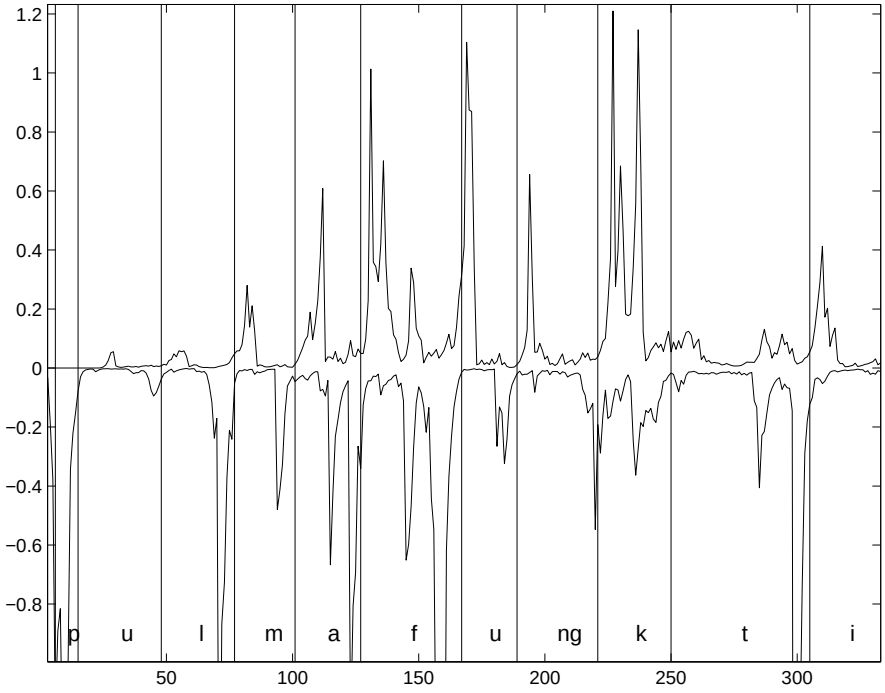

Figure 4.3: Forward and backwards prediction errors E_{t+} and E_{t-}, with manually segmented boundaries. E_{t-} has been given as negative values to help the visualization.

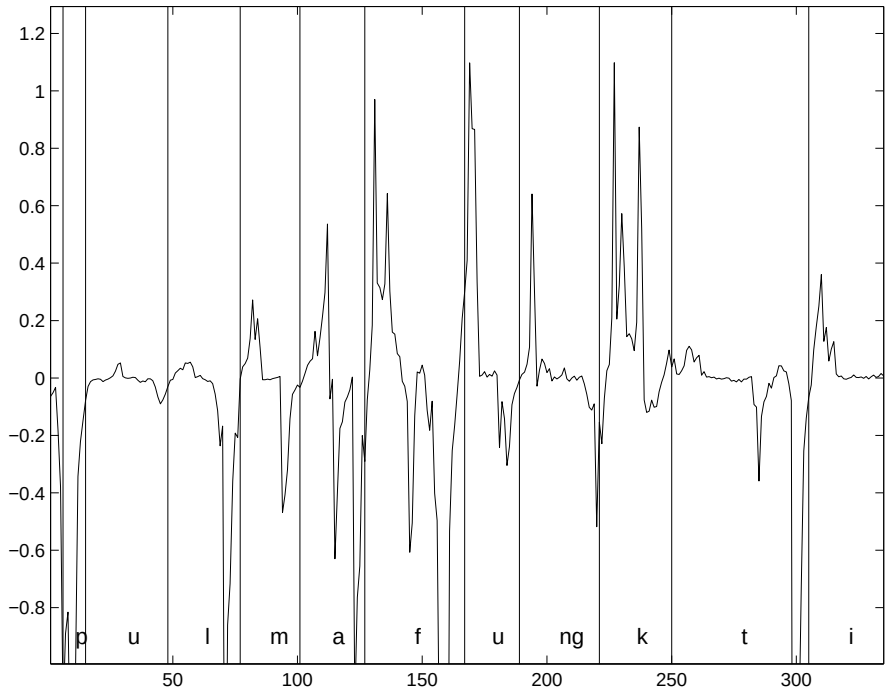

Figure 4.4: Combined backwards and forwards error signals e_{comb}

in Figure 4.4. There is a fair amount of local variation in the error E_{t*}, as we can see from the Figure 4.4. The signal is smoothed with a simple low-pass filter

$$H(z) = 1 + 0.95z^{-1} \tag{4.10}$$

The candidates for segment boundaries are now in the parts of signal, where the error goes rapidly to a large negative value, and shortly after that has a large positive value. In order to help the location of these parts, the signal E_{t*} is filtered with

$$h(t) = \begin{cases} \frac{t}{d} + 1 & -d < t < 0 \\ 0 & t = 0 \\ \frac{t}{d} - 1 & 0 < t < d \end{cases} \tag{4.11}$$

where d is set to be approximately the average width of peaks in the error. Filtering E_{t*} with $h(t)$ gives us a signal, where there are peaks at the points of segment boundary candidates. Local maxima can be detected by sliding a short window over the whole signal, and keeping track of the maximum values within the window. An example of the result we get from this can be seen in Figure 4.5.

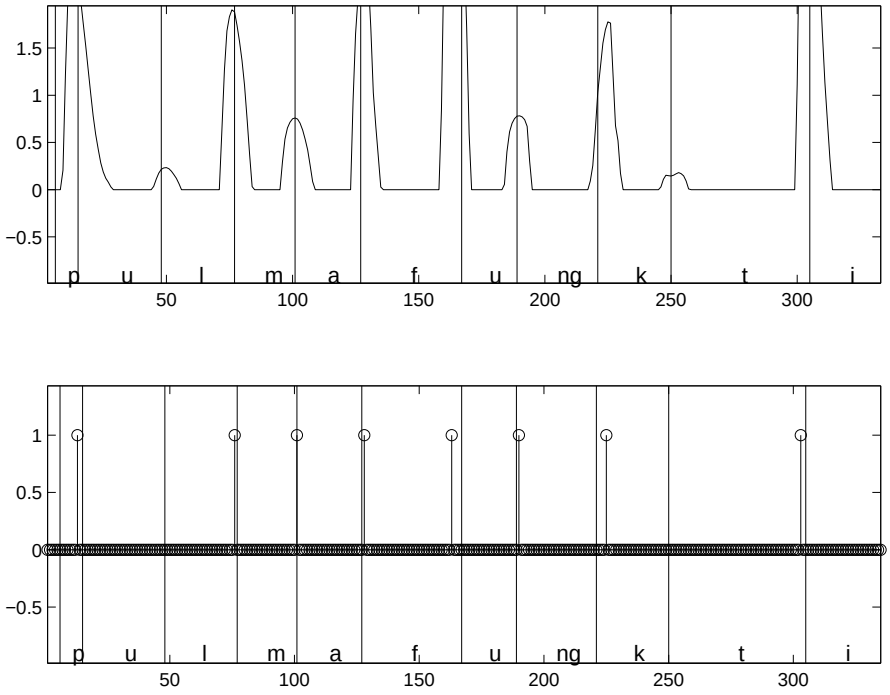

Figure 4.5: Up - error signal E_{t*} filtered with $h(t)$. Down - peak detection from the filtered error signal with a sliding window technique.

Chapter 5

Experiments

5.1 Performance Evaluation

5.1.1 Evaluation Criterion

Evaluation of performance of a segmentation algorithm is not straightforward. The method presented here detects acoustic landmarks, and it would be desirable that these acoustic landmarks would correspond to phonetic landmarks. Thus in this work a phonetic transcription was used to determine the identity of important acoustic-phonetic landmarks. Although there are many cases where it is difficult to precisely determine the location of a boundary between adjacent phones, the boundaries in the phonetic transcription usually corresponds to important landmarks which would be useful for subsequent acoustic-phonetic analysis of the speech signal. As such, they define a reasonable criterion for judging the performance of our algorithm. It is worth noting though, that differences between manual segmentation and automatic segmentation are not necessarily 'errors', especially if the differences occur in systematic manner. These differences might, for instance, point out inconsistencies in the nature of the phonetic transcription. At the present phase of the study, the ultimate goal is not to produce a comprehensive phonetic segmentation by detecting every phone boundary; we can concentrate especially to those boundaries which are the most reliable ones to detect. However, it may be expedient to see how far the basic method leads us.

The performance of the algorithm can be measured in various ways. In this work in addition to the obvious question of how accurately the method finds phoneme boundaries, we also investigate where the method performs poorly, and also how robust the system is under noisy conditions by gathering statistics on insertions and deletions. Also the performance under different parameter values is tested. The performance of the system at segment boundaries between different phoneme classes is investigated. All these statistics

together provide a reasonable indication of how well the system detects acoustic-phonetic information in the speech signal.

5.1.2 Performance Measures

To gain some useful knowledge about the performance of our system, we have to define some performance measures. There are two types of errors that the segmentation algorithm may introduce: deletion of segment boundaries, and insertion of boundaries where they do not belong. We shall denote total number of these errors with D (deletion), and I (insertion). These two types of errors are not unambiguous, because we have to define how far from the manual segmentation the automatic segmentation be placed for it to still be considered as correct segmentation. Depending on the application, we are interested in reducing some specific types of errors, and thus the analysis of different type of errors is valuable. Let us denote the total number of segment boundaries with N, and the number of correctly placed segment boundaries with H (hit). The *correctness* C (Eq. 5.1) describes the portion of segment boundaries placed correctly, and it is calculated by

$$C = \frac{H}{H + I} \tag{5.1}$$

Correctness is sometimes seen in the literature as *precision*. *Recall* R is the ratio of correctly placed boundaries to all manually placed segment boundaries calculated by

$$R = \frac{H}{N} \tag{5.2}$$

Quality Q (Eq. 5.3) is another performance measure. Quality takes the number of incorrectly inserted segment boundaries into account, and is calculated by

$$Q = \frac{H - I}{N} = \frac{N - D - I}{N} \tag{5.3}$$

The allowed deviation from manual segmentation for correct assignment of a segment boundary was set to $\pm 15ms$. This is justified noting, that the time resolution used in most ASR systems is $10ms$. In addition to these aforementioned performance measures, the temporal deviations from manually assigned segment boundaries were investigated.

5.1.3 Evaluation Data

The evaluation of the method was based on 201 Finnish language utterances combined from three different speakers (two males, and one female). The utterances were designed to cover as many different types of transitions between phoneme classes as possible. Talkers read the sentences aloud avoiding any emotional emphasis. Words were pronounced clearly and

more accurately than in normal conversations, though some of the test sentences were quite difficult to pronounce for native Finnish speakers, due to their somewhat artificial nature. Recordings were done in an anechoic room in the Laboratory of Acoustics and Audio Signal Processing at Helsinki University of Technology. The recordings were then digitized with 21.05 kHz sampling frequency. After the recording, the material was manually labeled with phonetic labels using Praat [1] software.

5.1.4 Preprocessing

Since the method described in this thesis is essentially a method of finding the acoustic landmarks which correspond to change in spectra of the speech, the digitized speech signal has to be converted from time domain into frequency domain representation. The representation chosen for this work was warped line spectrum pairs (WLSP), described in more detail in sections 3.1.2 and 3.1.3. The digitized speech signal was converted into warped LP coefficients of order 12, 14, and 16, and then these WLPC parameters were transformed into WLSP parameters. The time domain signal was windowed piecewise with a 20 ms Hamming window. There was a great amount of overlap between adjacent frames since the step size between frames was chosen to 3 ms. The unusually short step size between adjacent frames is appropriate for two reasons. Firstly, for the least squares estimation of the VAR model, the time series from where the model is estimated from has to be longer than the number of variables in a single vector. Secondly, the step size dictates the time resolution that the segmentation algorithm is capable of performing. This also matches the time resolution of the human hearing more accurately. In this work, the aim was to achieve good time resolution, and thus the step size selection which is different from the one used in standard ASR systems ($= 10ms$).

To investigate the robustness of the algorithm's performance under noisy conditions the original signal was corrupted with both pink and babble noise with different signal to noise (SNR) ratios before the analysis. Noise sequences were obtained from The Signal Processing Information Base (SPIB) [2].

[1] http://www.praat.org/. Praat is a tool for speech analysis developed by Paul Paul Boersma and David Weenink at Institute of Phonetic Sciences, University of Amsterdam

[2] http://spib.rice.edu/. The Signal Processing Information Base (SPIB) is a project sponsored by the Signal Processing Society and the National Science Foundation. SPIB is a repository of data, papers, software, newsgroups, bibliographies, and addresses as well as links to other relevant repositories.

5.2 Evaluation Results

5.2.1 Overall Results

In this section we look at the segmentation results for clean speech (not corrupted by noise). The threshold after which the deviation from the transcribed segment boundaries was considered as an error was set to 15 ms. First the overall results are shown, and the for each speaker separately. We shall call the speakers 'Male 1', 'Male 2', and 'Female'.

Table 5.2.1 summarizes the segmentation results of the three speakers. The overall quality Q did not vary significantly between the three speakers. The highest hit probability was achieved with the female voice, whereas the highest quality of performance was with the male speaker 1. The result confirms that the method is, by and large, speaker independent. It is worth noting that the quality (Q) is mostly reduced by the deletion rate (D), not the accuracy (C) of the segmentation. This suggests that the method is fairly reliable, yet not comprehensive. In section 5.2.3 we will look in detail which transitions between different broader phoneme classes cause the most deletions (D).

Table 5.1: Segmentation results [%] for three different speakers (C: Correctness, D: Deletions, Q: Quality, P: Precision), $M = 7$, $L = 66ms$, threshold = 0.2

	C	D	Q	P
Male 1	87.3	26.0	48.7	74.0
Male 2	88.2	32.6	43.7	67.5
Female	91.5	35.2	47.8	64.8

5.2.2 Effect of Parameter Selection

Next, we concentrate on detailed results obtained for the speaker 'Male 1' using different values for system parameters. Results for clean speech case are summarized in Figures 5.1 - 5.18. The results are plotted for different parameter values of p (WLSP order), the number of prediction errors included M and L, the length of the sequence from which the $VAR(1)$ model is estimated. Different threshold values for peak detection was also used. In all figures the three performance measures C, D, and Q are included.

An overall result that can be seen from all the figures is that as the correctness C increases the deletion rate D also always increases. On the other hand, the relation of these two performance measures does not stay the same. The selection of threshold is the most obvious

way to increase C. By selecting a larger threshold, the system detects only the biggest changes in the time-frequency domain, thus producing segmentation that is more syllabic than phonetic.

The number of correctly detected boundaries C is constantly higher when a larger amount of prediction errors (M) is included. However, increasing M also increases the number of deletions (D). This difference is most obvious when changing from value $M = 7$ to $M = 9$. This effect of changing M is the same for all the values of p.

The length of the data window L seems to have a similar effect. From the figures we can observe that as L increases, the number of correctly set segment boundaries C increases, but at the same time the more and more segment boundaries are ignored. This applies to all values of p.

When comparing results in terms of the WLSP order p, we can see that, for models that are estimated from long time series, the effect of p is very small. The only effect of increasing the WLSP model order p in this case is that the number of ignored segment boundaries decreases. For models that are estimated from shorter time series, the effect of variation of p is more evident.

In the case when $p = 16$ the number of correctly assigned segment boundaries decreases significantly (down to less than 80%), although the number of ignored segment boundaries also decreases. This is possibly because the number of vectors that the VAR model is estimated from is almost the same as the number of elements in each vector. The least squares estimation algorithm needs a multiple time series to estimate the model $\mathbf{A}$ with Equation 3.31, and if the number of elements in each vector of the time series is almost the same or less than the length of the time series, the model tries to model all the details of the time series, instead of capturing the trend of the original time series.

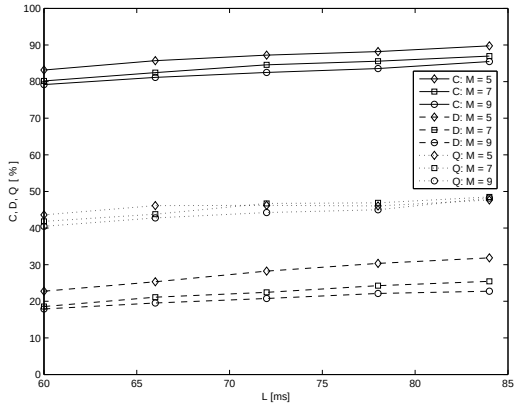

Figure 5.1: Segmentation accuracy p = 12, threshold = 0.10

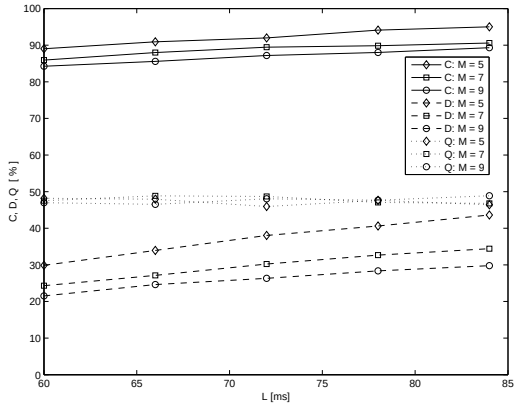

Figure 5.2: Segmentation accuracy p = 12, threshold = 0.20

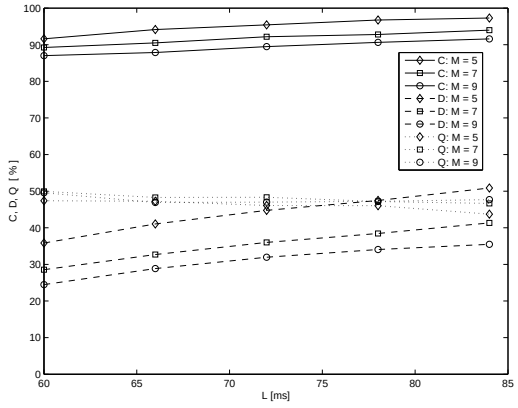

Figure 5.3: Segmentation accuracy p = 12, threshold = 0.30

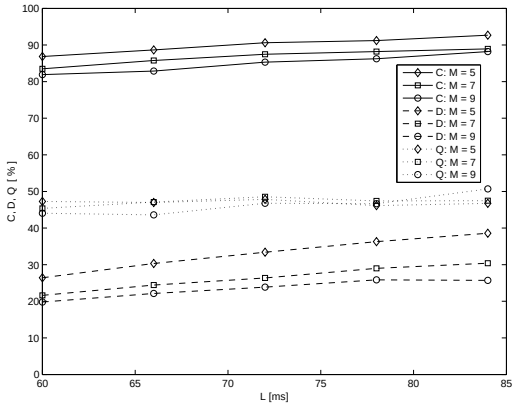

Figure 5.4: Segmentation accuracy p = 12, threshold = 0.15

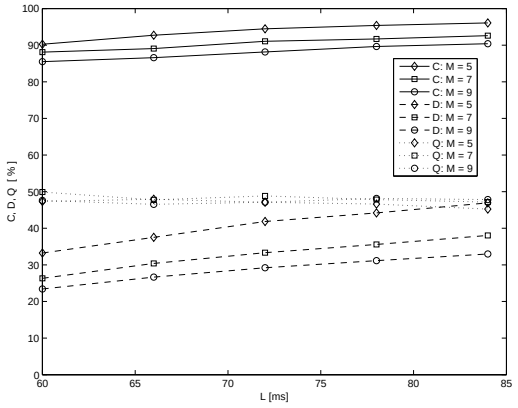

Figure 5.5: Segmentation accuracy p = 12, threshold = 0.25

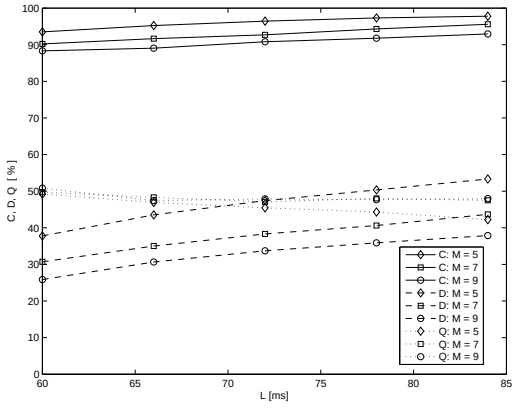

Figure 5.6: Segmentation accuracy p = 12, threshold = 0.35

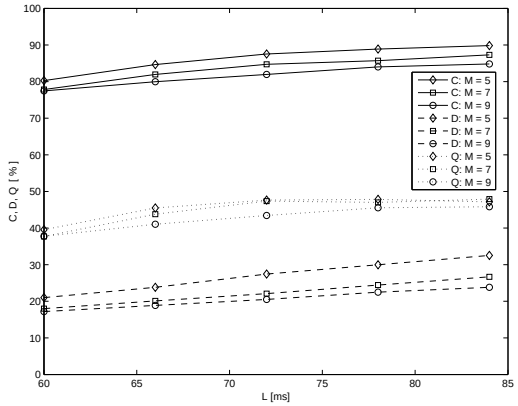

Figure 5.7: Segmentation accuracy p = 14, threshold = 0.10

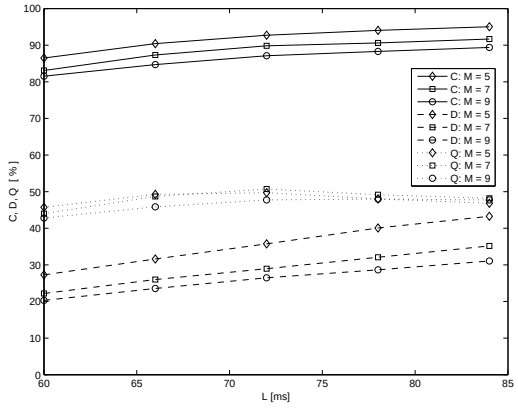

Figure 5.8: Segmentation accuracy p = 14, threshold = 0.20

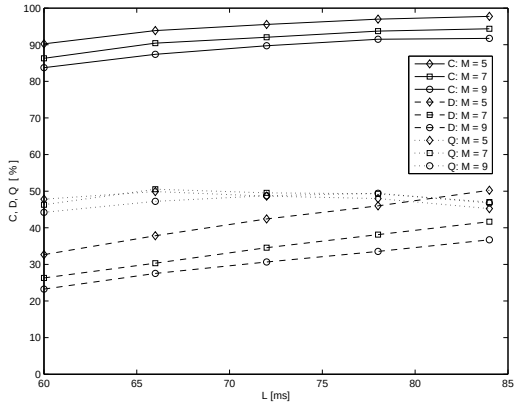

Figure 5.9: Segmentation accuracy p = 14, threshold = 0.30

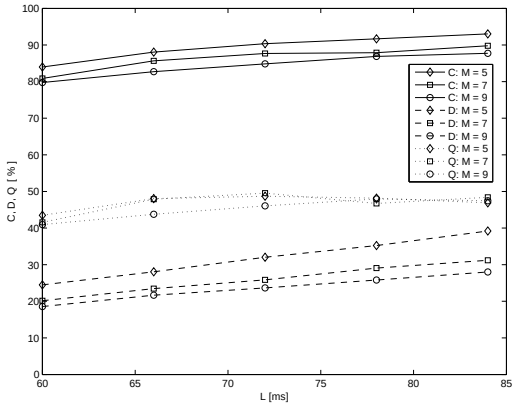

Figure 5.10: Segmentation accuracy p = 14, threshold = 0.15

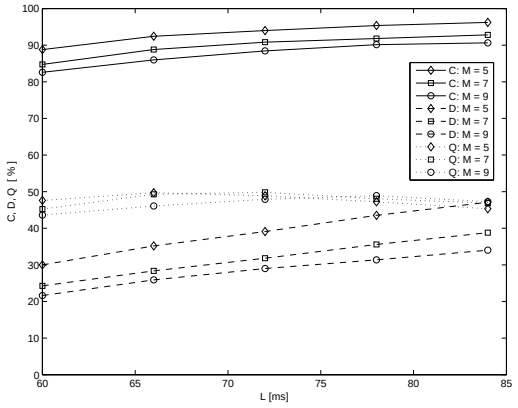

Figure 5.11: Segmentation accuracy p = 14, threshold = 0.25

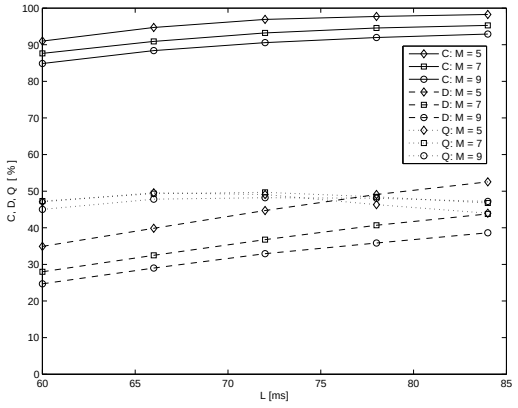

Figure 5.12: Segmentation accuracy p = 14, threshold = 0.35

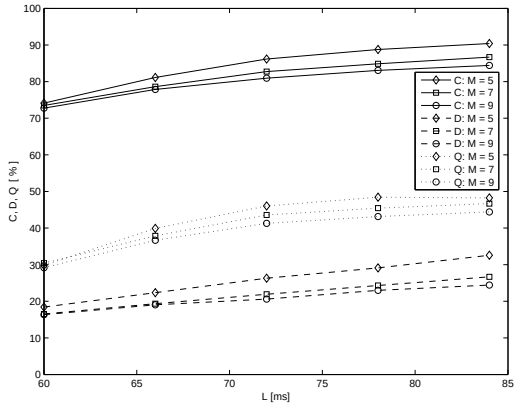

Figure 5.13: Segmentation accuracy p = 16, threshold = 0.10

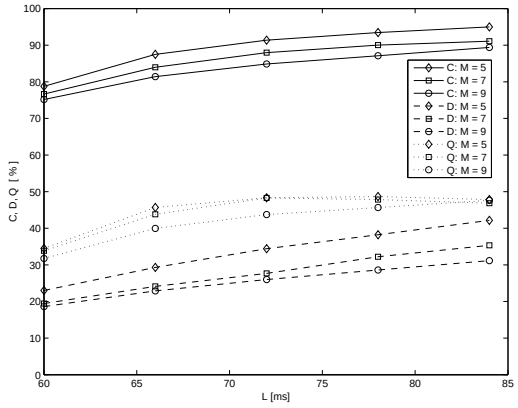

Figure 5.14: Segmentation accuracy p = 16, threshold = 0.20

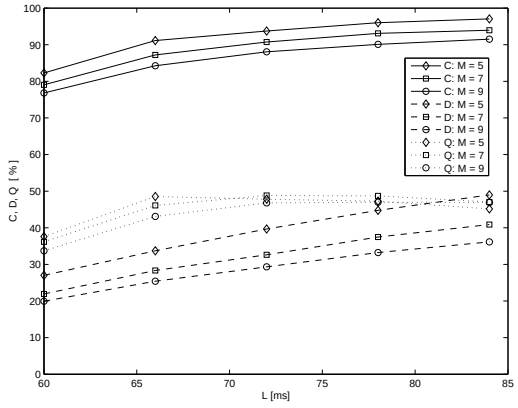

Figure 5.15: Segmentation accuracy p = 16, threshold = 0.30

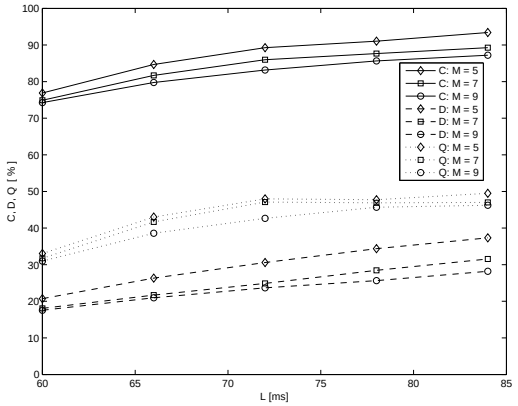

Figure 5.16: Segmentation accuracy p = 16, threshold = 0.15

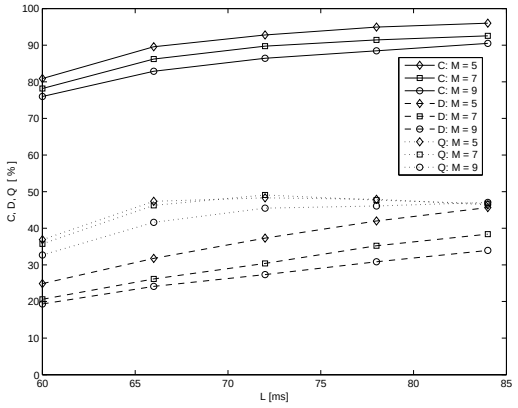

Figure 5.17: Segmentation accuracy p = 16, threshold = 0.25

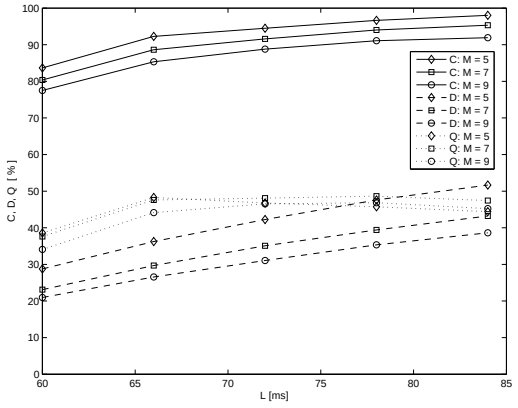

Figure 5.18: Segmentation accuracy p = 16, threshold = 0.35

5.2.3 Errors in Terms of Phoneme Classes

The way the acoustic signal changes at the segment boundary obviously depends on the phones around the segment boundary. To analyze the performance of our system thoroughly, we should investigate how the system works in all the different transitions from one phoneme to another. This, though, would be tedious and require more test data and time than was available for this analysis. Nevertheless, we can gain valuable information by grouping phonemes into some subclasses. The grouping scheme selected for this analysis was based on articulatory phonetics. In Section 2.2.2, the different classes of phonemes was shortly reviewed. This grouping scheme can be justified, if we remember that the phonemes belonging to a same class share characteristics in the way the are pronounced, and the manner of articulation reflects the properties of the acoustic signal. Laterals and semi-vowels were grouped together, because only a few instances of these phonemes were included in the test material.

The statistics were collected from segmentation results for speaker 'Male 1' using following parameter values: $M = 7$, $L = 66ms$, threshold $= 0.2$, $p = 14$. From Figure 5.8, we can see that the performance of the system is good with this selection of parameter values.

Theoretically, there are 49 different kind of transitions between the 7 phoneme classes. Seven of these transitions were not present in the material. Five of them are not realizable at all, or there are conflicting phonological rules of Finnish (marked with $\times$). Two of the cases are possible, but were not present in the material (marked with 0). 34 of 42 cases have three or more occurances.

The results for this test are shown in Table 5.2 with C and the number of occurances of each segment types n. The results show that the easiest segment boundaries to detect are vowel-stop and stop-vowel transitions which were detected with 94.0-96.5% accuracy. The most problematic segment boundries are the boundaries between two vowels. These are detected with 18.2% accuracy. Segment boundaries involving fricatives are detected with high accuracy. For segment boundary types which only occur less than three times, the analysis does not bring any information to light.

The last row of Table 5.2 shows the number of insertions in each phoneme class. The most common class for insertions is vowels, but vowels are also the most common class. If the number of insertions is compared with the relative probability of each segment, we can see that the most insertions occur in nasals, laterals+semivowels, trills, and silences.

Table 5.2: Percentage of correctly detected segment boundaries between phoneme classes, and number of insertions. Male speaker, $M = 7$, $L = 66ms$, $p = 14$ threshold = 0.2. Total number of segments 2264.

	vowels	stops	nasals	fricatives	laterals + semivowels	trill	silence
vowels	18.2	94.0	71.1	90.8	57.2	85.7	66.7
$n =$	181	300	190	141	138	70	12
stops	96.5	54.5	85.7	95.5	100	80.0	
$n =$	374	11	7	22	8	15	0
nasals	87.6	92.1	50.0	88.2	38.5	80.0	100
$n =$	121	38	4	17	13	5	4
fricatives	87.7	92.3	100	0	69.2	75.0	
$n =$	162	26	3	2	13	4	0
laterals + semi-vowels	51.5	87.5	100	100	20.0		
$n =$	165	8	2	9	5	$\times$	$\times$
trill	26.3	69.2	50.0	75.0	0		
$n =$	76	13	2	4	3	$\times$	$\times$
silence	75.0	30.2	75.0	37.5	88.9	75.0	
$n =$	12	43	12	16	9	4	$\times$
insertions	145	16	33	5	8	17	19
insertion [%]	13.3	3.6	15.0	2.4	4.2	17.4	19.8
$n =$	1091	439	220	211	189	98	96

5.2.4 Amount of E_* at Segment Boundaries

Spectral change at different phoneme class boundaries is not similar between classes, thus it is interesting to investigate what is the amount of error E_* at these points. The value of error at the boundary might give us some indication what kind of segment boundary is in question. An analysis of the E_* at manually assigned segment boundaries was conducted for speaker 'Male 1' using parameter values $p = 14$, $M = 7$, $L = 66ms$. The local maximum around the manually assigned segment boundary was located for all segment boundaries. Distributions of the values of E_* were computed, and are presented in Figure 5.19. For vowel-vowel transitions, the amount of error is low, which was expected since this segment boundary is the most difficult to detect using threshold. The same effect, even though not so dramatic, is between vowel and 'lateral+semivowel' classes. For vowel-stop and stop-vowel transitions, the values of E_* are concentrated well above zero. Distributions of E_* in transitions involving fricatives do not have distinct peaks, but values of E_* are spread evenly. Nasal-fricative transition has two peak distribution. This could be explained by remembering that the phoneme /h/ is sometimes slightly voiced, and in transition from voiced /h/ to nasal there is no radical spectral change.

These E_* distributions between phoneme classes could be used as a discrete transition probability distributions in pronunciation network in segmental speech recognition.

5.2.5 Temporal Deviation from the Manually Assigned Segment Boundaries

"Blind" segmentation algorithms such as the one introduced in this thesis are speech analysis tools that could be used in multitude of applications. Depending on the application, different performance measures are of interest. For example, it was emphasized in Section 1.1.3 that temporal alignment of speech signals for recognition of stop consonants has been overlooked in mainstream speech recognition. If we wish to use our segmentation algorithm as a front-end for segmental speech recognition system, we are interested of the temporal accuracy of the system.

Distribution of deviations from the manually assigned segment boundaries was collected for speaker 'Male 1' for clean speech. Since the segmentation algorithm works in symmetric fashion (forward-backward prediction for vector autoregression), the distributions were symmetric, and the data could be collected without an emphasis if the automatically set segment boundary was before or after the manually set boundary.

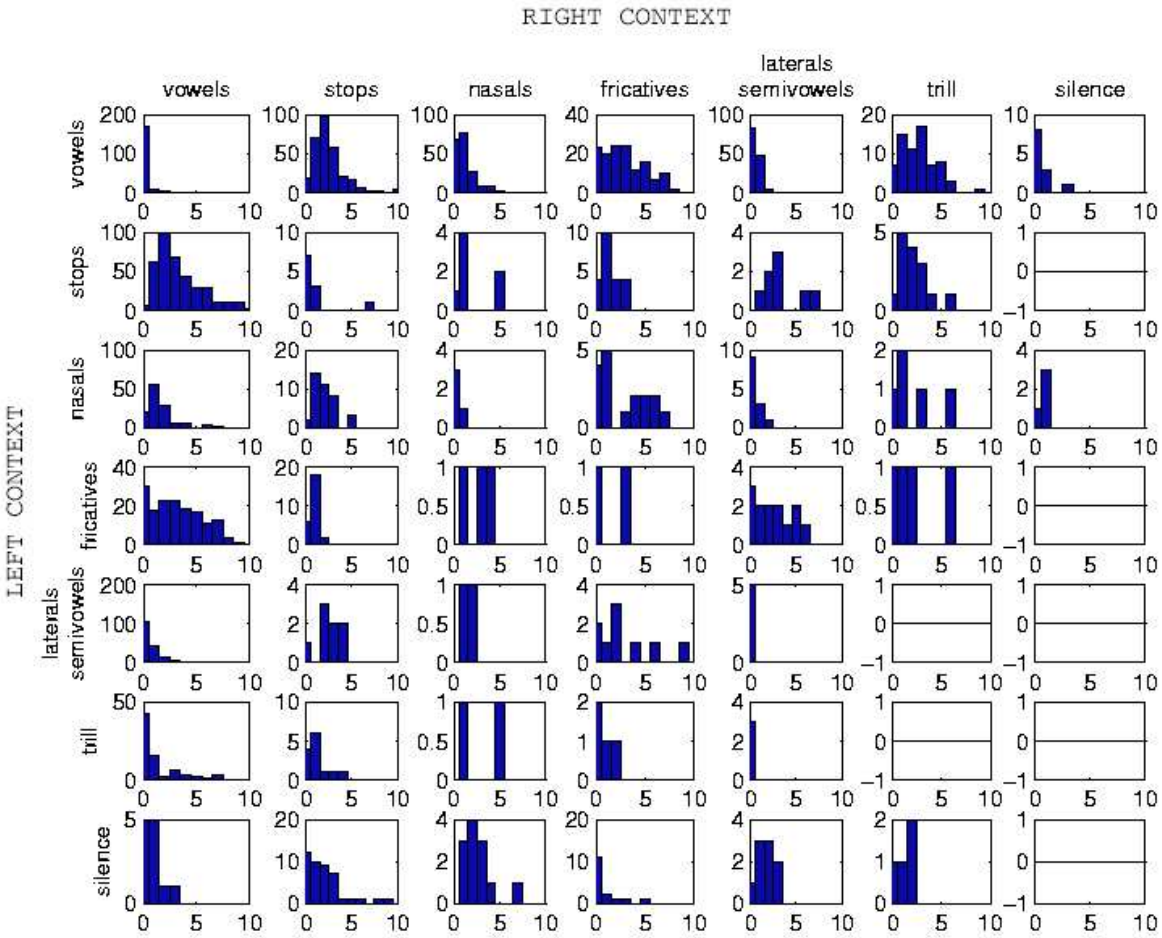

Figure 5.19: Histograms of amounts of errors around manually located phoneme class boundaries. Parameter values used: $p = 14$, $M = 7$, $L = 66ms$.

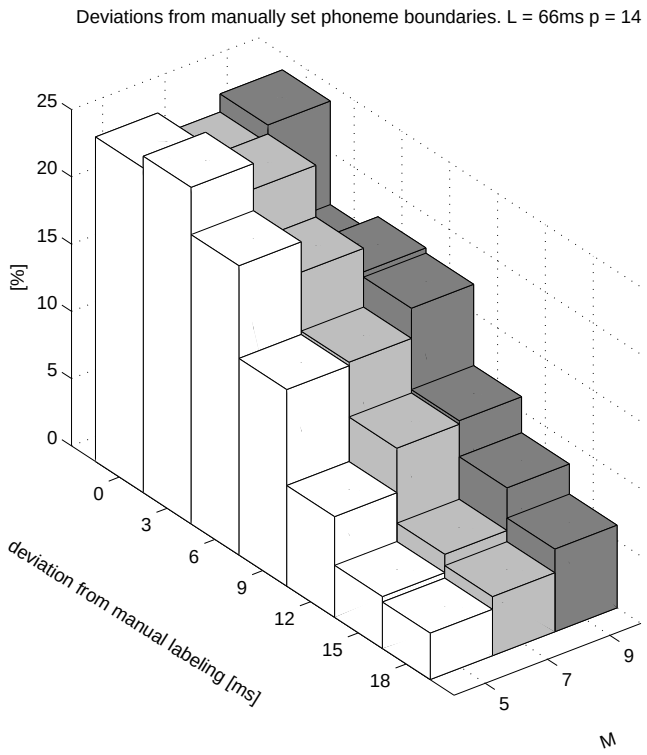

Figure 5.20: Deviation from the manually labeled boundaries using three different values of M.

In Figure 5.20 an example of three different distributions of temporal deviations is shown. The deviations are in each case concentrated around $0 - 3ms$. It should be noted that for each case $M = 5, 7, 9$, the deletion D and correctness C is higher and thus comparison between these three cases is not possible. The mean and the variance of the temporal deviations for the nine cases are listed in Table 5.3. Both mean and variance increase a little when M and threshold are increased respectively. It is also worth noting here that the performance measures C and Q are different in each case, thereby making the comparison of these means and variances difficult.

5.2.6 Effect of Noise

The segmentation performance was tested by controlling the sentence level SNR by adding both pink and babble noise to the speech signal. The results are presented in Fig. 5.21 with parameter values $L = 66ms$, $p = 14$, threshold = 0.2. The quality of performance drops

Table 5.3: Means and variances of the deviations from the manually asigned segment boundaries ($L = 66ms$, $p = 14$).

	M = 5	M = 7	M = 9
threshold = 0.1	mean: 1.8 ms variance: 5.9 ms	mean: 1.9 ms variance: 7.0 ms	mean: 2.2 ms variance: 8.5 ms
threshold = 0.2	mean: 1.8 ms variance: 6.0 ms	mean: 2.0 ms variance: 7.2 ms	mean: 2.3 ms variance: 8.8 ms
threshold = 0.3	mean: 1.9 ms variance: 6.0 ms	mean: 2.1 ms variance: 7.4 ms	mean: 2.3 ms variance: 9.0 ms

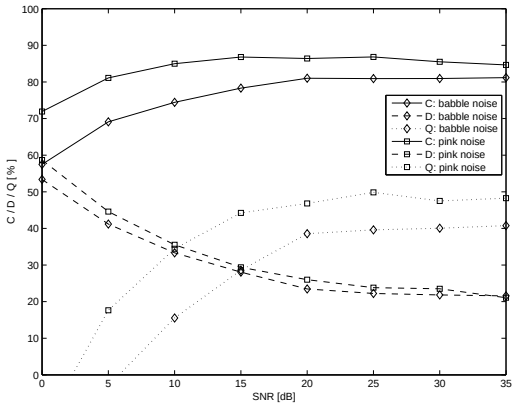

Figure 5.21: Segmentation accuracy in noisy conditions $p = 14$, threshold $= 0.20, M = 7, L = 66ms$

significantly when SNR goes below 15dB. The additive babble noise decreased the quality faster than the pink noise. This is intuitive, since the method is trying to detect changes in the speech spectrum, and the pink noise signal spectrum does not vary over time, whereas the babble noise spectrum is constantly changing.

5.2.7 Computational Load

The execution time of the algorithm was also shortly investigated. The system under which the algoritm was tested was modern standard PC (2 GHz Intel® Pentium® 4 processor, 512 MB of RAM) running Linux. All the algorithms were written for Matlab® software without any special optimization. The computational load depends on the selection of parameters such as order of the linear prediction, length of the sequence from which the VAR model is estimated, and temporal resolution of the segmentation (i.e. the step size for computation of LP). Since this algorithm was not a part of any particular system, but instead an attempt to investigate the possibilities to utilize vector autoregressive modeling for this

specific speech analysis subtask, the computational load was not considered a major issue. Thus the computational load is examined here only in the form of a single test with some typical parameter values.

In the test the selected parameters were the 14-th order WLSP ($p = 14$) computed in 20ms window with a 3ms step size. VAR models were estimated from 24 vectors ($L = 72ms$), and each model was used to iteratively produce 7 vectors ($M = 7$). With this setup the system execution time was 1.955 times the realtime.

Chapter 6

Conclusion and Perspectives

In this thesis a novel method to detect unpredictable auditory time-frequency changes in acoustic signals was introduced. The method is based on VAR-modeling of auditory spectrograms, and thus does not apply any *a priori* knowledge of the signals chosen for segmentation. Therefore the method is fully unsupervised and immediately applicable without any prior training. An interesting property of the method is that it matches the human auditory system in many respects, and is fully signal independent. This property allows its use in other fields of audio signal processing as well.

The method was tested on three speakers for the Finnish language. The results show that the introduced method is, by and large, speaker independent. The segmentation is reliable between classes that produce abrupt spectral changes at segment boundaries. Segment boundaries between vowel-vowel pairs proved the most difficult to detect.

6.1 Future Work

The performance of the method was measured in terms of its correlation to manual phonetic transcription. This can be considered as a reasonable benchmark for any segmentation system, but by no means is it ubiquitous. When testing the segmentation systems this way there are factors that affect the performance: for example, selection of the phoneme set, the way the manual segment boundaries are assigned. The latter is a source of a systematic error that we can not rectify. The former problem is of variable interest depending on our application.

In this work the multivariate time series used for VAR modeling consisted of warped line spectrum frequencies. This selection of variables was not proved to be optimal, and in the

future our aim is to investigate the use of other frame based parameters of speech signal for VAR based segmentation. There exists a number of different features that have been suggested to help categorize the speech signal into meaningful classes. These same features might be used to find segment boundaries between two different types of speech segments. These features include for example: log-energy, zero-crossing rate, energy across different frequency bands, autocorrelation coefficient at unit sample, normalized prediction error, and so forth. [39]. In the VAR method presented here, the multivariate time series can consist of vectors whose elements are scaled differently, and thus we can combine several of these parameters to our vector representation of the signal. Systematic analysis of different parameters at the phone boundaries could give us an optimal set of features to use for segmentation.

The tests we conducted in this thesis produced an explicit segmentation of the speech signal (single level of segmentation). This type of segmentation scheme always produces errors in the form of deletions and substitutions, which might be crucial for some applications. One possible method to overcome this is to build a multi-level segmentation [40]. For our algorithm this could be done using different decision rules to select the boundaries from the forward-backward prediction error. With this kind of approach, instead of the single-segmantation scheme for the utterance, we would have a dendogram of different possible segmentations for the utterance. This kind of approach could be used for example in segment-based speech recognition.

Even though the presented segmentation method is signal independent, the nature of the test material itself might have an effect on the results. The method should not be language depepndent, but as shown in this work, deletions occur more often between some phonetic boundaries than others; thus, the types of typical phoneme transitions in the language affects the performance. The set of test utterances used in this work was constructed to cover as many different transitions as possible in a compact set. The speech consisted of many occurances of sequence of several vowels (e.g. /lieoissa/), something that is not present in many languages. In the near future, the method will be tested with English-language material obtained from TIMIT speech database.

Bibliography

[1] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, Inc., 1993.

[2] Jane W. Chang. *Near-Miss Modeling: A Segment-Based Approach to Speech Recognition*. Doctoral thesis, Massachusetts Institute of Technology, 1998.

[3] Markku Ursin. Triphone clustering in finnish continuous speech recognition. Master's thesis, Helsinki University of Technology, Department of Computer Science, Laboratory of Computer and Information Science, 2002.

[4] Richard P. Lippman. Speech recognition by machines and humans. *Speech Communication*, 22:1–15, 1997.

[5] Amit Juneja. *Speech Recognition using Acoustic Landmarks and Binary Phonetic Feature Classifiers*. Doctoral thesis (proposal), University of Maryland, 2003.

[6] H. Flecher and J. C. Steinberg. Articulation testing methods. *Bell Syst. Tech. J.*, 88:809–854, October 1929.

[7] Jont B. Allen. How do humans process and recognize speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):567–577, October 1994.

[8] James R. Glass. A probabilistic framework for segment-based speech recognition. *Computer Speech and Language*, 17:137–152, 2003.

[9] Toni Hirvonen and Unto K. Laine. Comparison of objective and subjective classification of unvoiced stop consonants in stop-vowel syllables. *Proceedings of 3rd International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA2003)*, pages 265–268, December 2003.

[10] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon. *Spoken Language Processing*. Prentice Hall, 2001.

[11] Kenneth N. Stevens. *Acoustic Phonetics*. The MIT Press, 1998.

[12] Douglas O'Shaugnessy. *Speech Communication*. Series in Electrical Engineering. Addison-Wessley, 1987.

[13] Lade. *Course in phon.* Brooks/Cole Publishing Company, 1993.

[14] Thomas D. Rossing. *The Science of Sound*. Addison-Wesley Publishing Company, 2 edition, 1990.

[15] E. Zwicker and E. Ternhart. Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *J. Acoust. Soc. Am*, 68:1523–1525, 1980.

[16] Richard M. Warren. *Auditory Perception*. Cabridge University Press, 1999.

[17] B. Lindblom and M. Studdert-Kennedy. On the role of formant transitions in vowel recognition. *Journal of the Acoustical Society of America*, 42:830–843, 1967.

[18] Ronald A. Cole, editor. *Perception and Production of Fluent Speech*. Lawrence Erlbaum Associates, Publishers, 1980.

[19] John Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, April 1975.

[20] F. Itakura. Line spectrum representation of linear predictor coefficients of speech signals. *Journal of Acoustical Society of America*, 57:S35, April 1975.

[21] Hans Werner Strube. Linear prediction on a warped frequency scale. *Journal of Acoustical Society of America*, 68:1071–1076, October 1980.

[22] J. Makhoul and M. Berouti. Adaptive noise spectral shaping and entropy coding in predictive coding of speech. *IEEE Trans. Acoust., Speech, Signal Process.*, ASSP-27:63–73, February 1979.

[23] Aki Härmä, Matti Karjalainen, Lauri Savioja, Vesa Välimäki, Unto K. Laine, and Jyri Huopaniemi. Frequency-warped signal processing for audio applications. *Journal of Audio Engineering Society*, 48:1011–1031, November 2000.

[24] Aki Härmä. Implementation of frequency-warped recursive filters. *Signal Processing*, 80:543–548, February 2000.

[25] Ivan Margin-Chagnolleau, Joachim Wilke, and Frédéric Bimbot. A further investigation on ar-vector models for text-independent speaker identification. *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, 1:101–104, May 1996.

[26] F. Bimbot, L. Mathan, A. De Lima, and G. Chollet. Standard and target driven ar-vector models for speech analysis and speaker recognition. *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, 2:5–8, March 1992.

[27] Claude Montacié and Jean-Luc Le Floch. Ar-vector models for free-text speaker recognition. *Proceedings of ICSLP 92*, 1:611–614, October 1992.

[28] Claude Montacié, Paul Deléglise, Frédéric Bimbot, and Marie-José Caraty. Cinematic techniques for speech processing: Temporal decomposition and multivariate linear prediction. *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, 1:153–156, March 1992.

[29] Helmut Lutkepohl. *Introduction to Multiple Time Series Analysis*. The McGraw-Hill Companies, Inc., 2 edition, 1993.

[30] L.R. Rabiner, A. E. Rosenberg, J. G. Wilpon, and T. M. Zampini. A bootstrapping training technique for obtaining demisyllabic reference patterns. *J. Acoust. Soc. Amer.*, 71:1588–1595, 1982.

[31] J.G. Wilpon, B.H. Juang, and L.R. Rabiner. An investigation on the use of acoustic sub-word units for automatic speech recognition. *Proc. of IEEE Internat. Conf. on Acoust., Speech and Signal Processing*, pages 821–824, 1987.

[32] Jan P. H. van Santen and Richard W. Sproat. High-accuracy automatic segmentation. *Proc. Eurospeech*, 6:2809–2812, 1999.

[33] A. Vorstermans, J.P. Martens, and B Van Coile. Automatic segmentation and labeling of multi-lingual speech data. *Speech Communication*, 19:271–293, 1996.

[34] Kris Demuynck and Tom Laureys. A comparison of different approaches to automatic speech segmentation. *Proceedings of the 5th International Conference on Text, Speech and Dialogue*, pages 277–284, 2002.

[35] P. Mermelstein. Automatic segmentation of speech into syllabic units. *Proceedings of ICSLP 92*, 1:611–614, October 1992.

[36] Manish Sharma and Richard Mammone. ”blind” speech segmentation: automatic segmentation of speech without linguistic knowledge. *Proceedings., Fourth International Conference on Spoken Language*, 2:1237–1270, October 1996.

[37] Guido Aversano, Anna Esposito, Antonietta Esposito, and Maria Marinaro. A new text-independent method for phoneme segmentation. *MWSCAS 2001 Proceedings of the 44th IEEE 2001 Midwest Symposium on Circuits and Systems*, 2:516–519, August 2001.

[38] V. Kamakshi Prasad, T. Nagarajan, and Hema A. Murthy. Automatic segmentation of continuous speech using minimum phase group delay functions. *Speech Communications*, 42:429–446, 2004.

[39] Bishnu S. Atal and Lawrence R. Rabiner. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(3):201–212, June 1976.

[40] James Robert Glass. *Finding Acoustic Regularities in Speech: Applications to Phonetic Recognition*. Doctoral thesis, Massachusetts Institute of Technology, 1988.