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

Jouni Pohjalainen

Frequency-Warped Linear Prediction and Speech Analysis

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

Espoo, Mar 12, 2004

Supervisor: Professor Unto K. Laine

HELSINKI UNIVERSITY OF TECHNOLOGY	ABSTRACT OF THE MASTER'S THESIS

Author:	Jouni Pohjalainen		
Name of the thesis:	Frequency-Warped Linear Prediction and Speech Analysis		
Date:	Mar 12, 2004	**Number of pages:**	72
Department:	Electrical and Communications Engineering		
Professorship:	S-89		
Supervisor:	Prof. Unto K. Laine		

Linear prediction is a popular method in digital signal processing for spectrum envelope estimation. Models for the spectrum envelope are useful in e.g. speech analysis tasks such as feature generation for automatic speech recognition.

The frequency representation used by ordinary linear prediction can be transformed by frequency-warping techniques to approximate e.g. the auditory Bark scale. Using frequency-warped linear prediction, it is thus possible to model a wideband spectrum with a frequency resolution close to that of the human auditory system. This way, fewer predictor coefficients can be used to model the auditorily most relevant information.

This Master's thesis studies warped linear prediction techniques with the emphasis on modeling the spectrum of speech. Objective model quality measures have been developed and applied to the study of the main differences between ordinary and Bark-warped linear prediction. The optimal reduced prediction order, or number of model parameters, in warped linear prediction of speech has been determined using the modeling performance measures. In addition, two frame-based and two computationally more efficient adaptive methods have been analyzed and compared with each other. Results are reported on how the time resolution of these analysis methods can be adjusted properly.

Keywords: Frequency warping, linear prediction, speech analysis

TEKNILLINEN KORKEAKOULU DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Jouni Pohjalainen		
Työn nimi:	Taajuusalueessa varpattu lineaarinen ennustaminen ja puheen analyysi		
Päivämäärä:	12.3.2004	**Sivuja:**	72
Osasto:	Sähkö- ja tietoliikennetekniikka		
Professuuri:	S-89		
Työn valvoja:	Prof. Unto K. Laine		

Lineaarinen ennustaminen on digitaalisessa signaalinkäsittelyssä paljon käytetty menetelmä spektrin verhokäyrän estimointiin. Spektrin verhokäyrämallit ovat käyttökelpoisia esimerkiksi monissa puheanalyysisovelluksissa, kuten automaattisen puheentunnistuksen piirteiden muodostusvaiheessa.

Tavallisen lineaarisen ennustamisen käyttämä taajuusesitys voidaan muuntaa niin sanotuilla varppaustekniikoilla siten, että signaalinkäsittelyjärjestelmän taajuusresoluutio vastaa esimerkiksi ihmisen kuulon mukaista Bark-asteikkoa. Taajuusalueessa varpatulla lineaarisella ennustamisella voidaan täten tuottaa kuulon taajuusresoluutiota vastaavia spektrimalleja laajakaistaisista signaaleista. Menetelmän etuna on se, että auditorisesti olennainen informaatio voidaan kuvata pienemmällä määrällä malliparametreja.

Tämä diplomityö käsittelee taajuusvarpatun lineaarisen ennustamisen toteutuksia erityisesti puheen spektrin mallintamisen kannalta. Työssä on kehitetty objektiivisia mallien laatumittoja, joilla on tarkasteltu tavallisen ja Bark-asteikolle varpatun lineaarisen ennustamisen keskeisiä eroja. Bark-asteikolle varpatulle lineaariselle ennustamiselle on määritetty näiden mittojen avulla puhekommunikaation kannalta sopivin alennettu asteluku eli malliparametrien määrä. Lisäksi työssä on tutkittu ja vertailtu keskenään kahta kehyspohjaista ja kahta laskennallisesti hieman tehokkaampaa adaptiivista menetelmää. Työssä on selvitetty, miten eri menetelmien aikaresoluutio voidaan parhaiten asettaa halutuksi.

Avainsanat: Taajuusvarppaus, varppaus, lineaarinen ennustaminen, lineaariprediktio, puheanalyysi

Preface

This Master's thesis has been done in the Laboratory of Acoustics and Audio Signal Processing at the Helsinki University of Technology. The motivation for the work arose during the Usix-STT speech recognition project, financed by Tekes, in which I was working as a research assistant. The thesis work was continued, among other research activities, after the project ended in 2003.

I want to thank my thesis supervisor, professor Unto K. Laine for employing me and for sharing his expertise in various fields relating to speech technology and signal processing. I feel I have learned a lot. I also thank Dr. Aki Härmä for providing an implementation of the gradient adaptive lattice algorithm and for some helpful comments. Many other people at the laboratory have also offered interesting discussions, valuable insights into different subjects, and help with practical matters, for which they all deserve thanks.

Otaniemi, March 12, 2004

Jouni Pohjalainen

Contents

Abbreviations

AR	Autoregressive
ARMA	Autoregressive Moving Average
DSP	Digital Signal Processing
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
GAL	Gradient Adaptive Lattice
IIR	Infinite Impulse Response
IPA	International Phonetic Alphabet
LMS	Least-Mean-Squares
LP	Linear Prediction
LPC	Linear Predictive Coding
MA	Moving Average
MFCC	Mel-Frequency Cepstral Coefficients
PLP	Perceptual Linear Prediction
SPL	Sound Pressure Level
WLP	Warped Linear Prediction

List of Figures

List of Tables

Part I

Theory

Chapter 1

Introduction

Physically, human speech manifests itself as air pressure variations caused by movements of the speech production organs of the speaker. When computers or signal processors are used to process a speech signal, the pressure variation signal is first converted to an electric signal by a microphone. The continuous electric signal is then sampled at discrete time points and quantized to a sequence of discrete-valued samples. A digital speech signal is thus obtained and stored in the memory of the device.

Applications involving digital speech signal processing include speech analysis, recognition, coding, and synthesis. *Speech analysis* is concerned with processing techniques for extracting information from the speech waveform [2]. Speech can be analyzed just for analysis' sake, e.g. when doing research of phonetics or speech production modeling. Speech analysis is also one processing stage in speech recognition, coding, and synthesis. In these applications, capturing the envelope, or general shape, of the spectrum is particularly important. A popular method for parametric spectrum estimation is linear prediction (LP) [22].

The primary motivation of this study is to serve the *feature generation* stage in speech recognition, which is a pattern recognition problem. The numerous discrete speech samples need to be represented by a smaller number of features. The features are usually chosen to represent aspects of the short-time spectra. LP can be used in a speech recognition front-end to generate features that represent the spectrum envelope [28].

Maximum reduction of the number of features with minimum loss of discriminative information is very desirable in pattern recognition. This applies also to speech recognition. One approach to reduce the number of features further is to exploit knowledge of the human auditory perception and focus on the *perceptually* most relevant parts of the signal spectrum. One way to do this is by spectrum estimation via frequency-warped linear prediction, or simply warped linear prediction (WLP), which is discussed in this thesis. WLP is a generalization of ordinary linear prediction and allows the modeled signal frequencies to be mapped into a nonuniform resolution frequency scale, such as the psychoacoustic auditory Bark scale.

Frequency-warped digital signal processing techniques, including WLP, have been extensively studied at the Laboratory of Acoustics and Audio Signal Processing of the Helsinki University of Technology [11]. Many different WLP techniques have been developed, implemented, and applied especially to wideband audio coding [13] [11] and speech synthesis [19]. The application of WLP in speech recognition front-ends has thus far not received overwhelming attention. WLP-based feature extraction for speech recognition was initially suggested after promising results in stop-vowel classification tests [21]. Since then, front-ends based on WLP have compared favorably against established techniques, such as mel-frequency cepstral coefficients (MFCC), in spoken digit recognition experiments [34] but less favorably in phonemic continuous speech recognition [33]. In any case, WLP-based front-ends have achieved results comparable with the standard auditorily motivated methods such as MFCC and Perceptual Linear Prediction (PLP). All these techniques have used the modified autocorrelation method of WLP, henceforth called simply the autocorrelation method as in ordinary LP. While the autocorrelation method has been dominant, several alternative feasible techniques exist for WLP computation, including block (frame-based) methods as well as *adaptive* methods.

The purpose of this work is not to give a comprehensive treatment of all possible frequency-warped versions of the numerous techniques for linear predictive coding. Rather, the focus is on a few important issues that have perhaps not received sufficient treatment thus far. These issues are related both to WLP in general and specific computation techniques. This work tries to find answers in particular to the following questions: How would one objectively select the model order when the modeled signal is a speech signal? How should we select the adaptation rate in adaptive WLP techniques? How do these computationally efficient adaptive techniques compare against the standard block techniques with different types of speech input?

The organization of the thesis is as follows. The theoretical part includes chapters 2-4. Chapter 2 reviews some important basic concepts of speech production, acoustic representation of the speech signal, and hearing. Chapter 3 discusses LP and WLP analysis. Chapter 4 introduces the model quality measures used in the experimental part. The empirical part includes chapters 5-8. Chapter 5 gives an overview of the simulations setup. Chapter 6 contains the results of analyses related to model order selection and also discusses the differences between LP and WLP. Chapter 7 discusses setting the temporal resolution of the block and adaptive estimation techniques optimally and also compares them with each other in terms of model quality. Chapter 8 presents the conclusions.

Chapter 2

Speech communication at the physical level

2.1 Speech production mechanism

The human speech organs, depicted in figure 2.1, can be divided into three main groups: lungs, larynx, and vocal tract. While the term vocal tract is sometimes used to refer to the whole system of organs involved in speech production, for speech modeling purposes it is more practical to include only the supralaryngeal articulators in the definition. The excitation for speech can occur in the larynx or in many different places along the vocal tract, depending on the sound to be generated. Speech sounds can be classified according to their type of excitation. The four main types of excitation are: *voiced*, *unvoiced*, *mixed* voiced and unvoiced, and *silence*.

The primary function of the lungs, situated in the chest or *thorax* cavity, is breathing. Breathing is accomplished primarily by contraction and relaxation of the diaphragm at the bottom of the thorax. The lungs inspire and expire a tidal volume of air every 3-5 s at rest. Normal exhaling takes about 60% of the duration of each breathing cycle [27]. The lungs also have an important role in speech production, because they generate a pressure difference that is important in producing the excitation for speech sounds. Most sounds in all languages are *egressive*, that is, they are formed during expiration. For this reason, the inhale-exhale time ratio during speech differs from that during normal breathing and is typically about 1:10.

The excitation for the voiced speech sounds occurs in the larynx, where the *vocal folds* are located. The vocal folds are a pair of elastic structures of tendon, muscles, and mucous membranes, with posterior ends attached to the *arytenoid cartilages* and anterior ends attached to the *thyroid cartilage* (Adam's apple). The opening between the vocal folds is

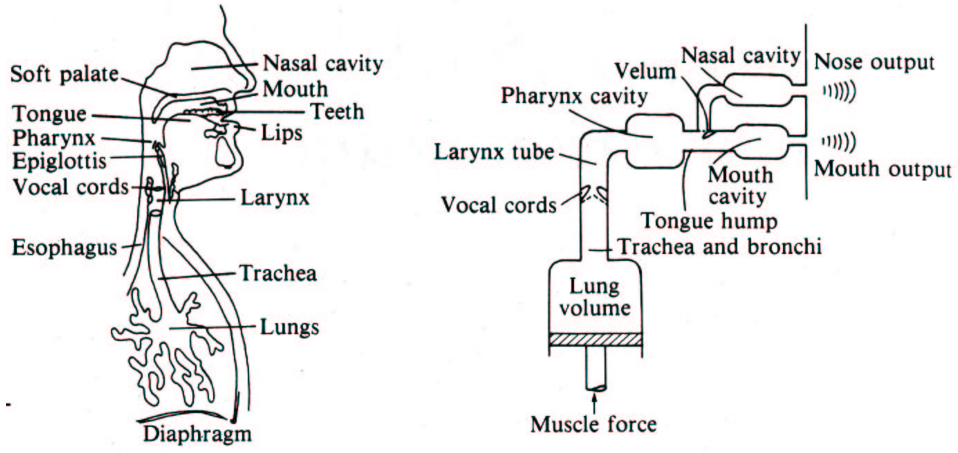

Figure 2.1: The human speech production mechanism (from [18]).

called *glottis*. During normal breathing, the glottis remains sufficiently open to allow free air passage without creating sound; however, when air is expelled from the lungs through the glottis and the tension of the vocal folds adjusted properly, the vocal folds can be made to vibrate in an oscillatory fashion. This causes periodic interruption of the subglottal airflow from the lungs, which in turn results in quasi-periodic pulses of air that excite the vocal tract.

Unvoiced excitation can occur either in the larynx or various parts of the vocal tract. *Whisper* (*aspiration*) sounds are generated when the glottis is made to stay narrow enough to hinder airflow from the lungs, thus creating turbulent noise. Noise can also be generated by the same mechanism somewhere in the vocal tract. Turbulent noise sounds generated at a narrow constriction in the vocal tract are called *fricatives*. Voiced and unvoiced excitation can sometimes be mixed, when both vocal fold vibrations and turbulent noise at a constriction are present simultaneously.

According to the *source-filter model* of speech production [4], the vocal tract acts as a filter shaping the signal from the excitation source (either voiced or unvoiced). The vocal tract resonances shape the spectrum envelope of the source. Speech sounds with similar excitation have their main acoustic differences in the spectrum envelopes, which are imposed by different vocal tract filters.

2.2 Acoustic phonetics

The science of *phonetics* studies issues related to speech sounds, their production, and perception. It can be roughly divided into *articulatory*, *acoustic*, and *auditive* phonetics [37].

Articulatory phonetics relates linguistic features of sounds to positions and movements of the speech organs. Acoustic phonetics studies speech waveforms and spectra and their relationships to phonemes (the basic representatives of different speech sounds). It deals with the differentiation of speech sounds based on the observed acoustic output of the speech production system. Auditive phonetics is concerned with speech perception. The emphasis in this thesis is on the acoustic aspects, as the main goal is to find methods for generating good spectrum envelope models and good modeling of *formants*.

The original primary meaning of the term formant is an observable spectral peak in the sound spectrum [4]. Since the spectral peaks are very closely linked with *resonance frequencies* of the vocal tract filter, these terms may often be used synonymously. In most applications, formant frequencies are taken to be properties of the vocal tract system during voiced phonation. Formants may be abbreviated Fi, where the first formant $F1$ is the formant with the lowest frequency. The low frequency formants, in particular the second formant $F2$, are the most important ones in discriminating between voiced speech sounds. Formants $F1$-$F3$ fall below 3500 Hz [28] and in most cases below 3000 Hz [27].

Spectrograms are a basic tool in spectral analysis. Instead of a common two-dimensional speech waveform (amplitude vs. time), a spectrogram is a three-dimensional pattern: a time sequence of short-time spectra (amplitude vs. frequency). An example spectrogram is shown in figure 2.2. With time and frequency on the horizontal and vertical axes, respectively, amplitude is noted by the darkness of the display. The dark bands with mostly horizontal orientation represent the formants.

The following phoneme categories are relevant in this thesis:

- *Vowels* are voiced (except when whispered), have great intensity, and are fairly long in duration (up to a few hundred milliseconds). The transitions between different vowels are gradual rather than abrupt due to the fact that the vocal tract shape can not be changed very fast.

- *Nasals* and *laterals* somewhat resemble vowels but are voiced, steady consonants during which the mouth output of the vocal tract is limited; in nasals, the primary vocal tract is closed and the air flows through the nasal cavity (figure 2.1); in laterals, the tongue blocks the main air flow by pressing against the alveolar ridge behind the teeth, but there is a passage for air flow on both sides of the tongue.

- *Semivowels* (glides) and *trills* are voiced consonants which are often less stationary than vowels, nasals, or laterals; there may be transient changes in the vocal tract and considerable constriction reducing intensity (semivowels) or modulation caused by tongue tip vibration (trills).

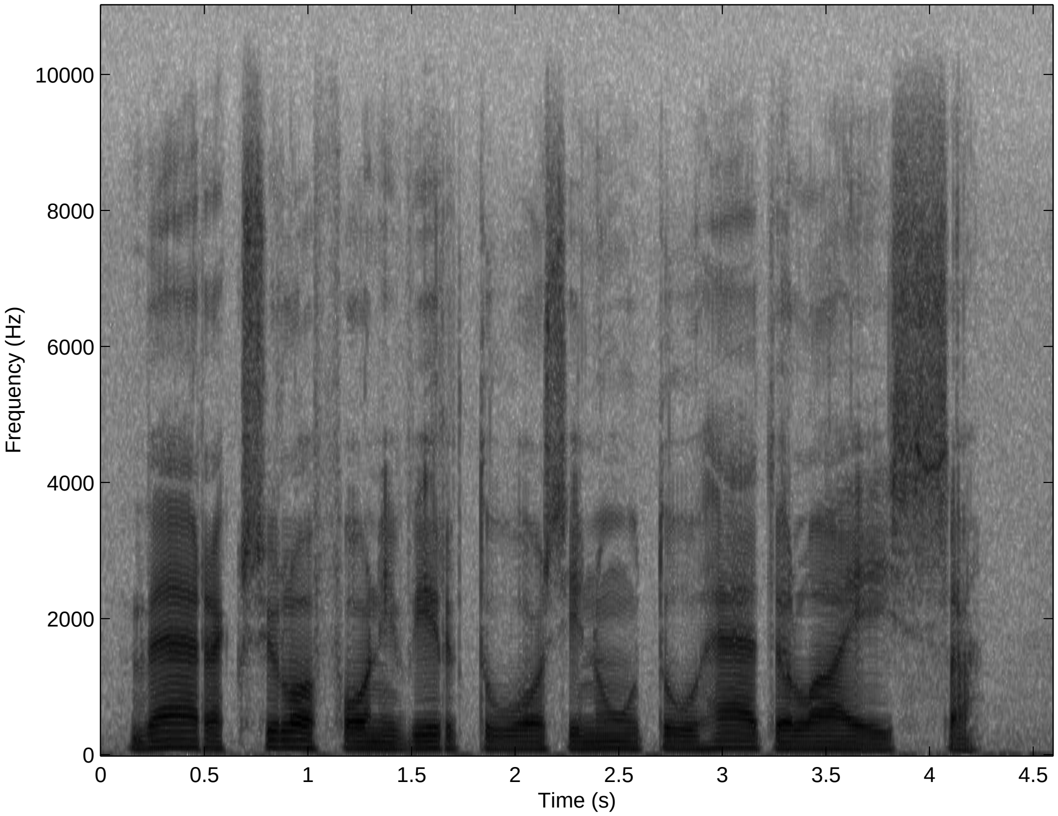

Figure 2.2: Example spectrogram of the utterance ’väärä ksylofonivirtuoosi nuutuu häätölaeissa’.

- *Fricatives* are usually unvoiced, aperiodic and with most part of the energy at high frequencies.
- *Stop consonants* or plosives are of transient nature and consist of two parts: the occlusion, or vocal tract closure, which renders the speech either silent (for unvoiced stops) or a low intensity low frequency voiced sound (voiced stops); the closure is ended by a noise burst usually followed by frication.

When it is necessary to refer to individual phonemes, this thesis uses the *fenno-ugric* system of phonemic transcription [37]. The notation in the present scope also complies with the Finnish orthography; the phonemes in spoken Finnish and the characters in written Finnish have an almost one-to-one relation. Most of the characters in written Finnish are identical to the corresponding phoneme symbols in the International Phonetic Alphabet (IPA) with only a few exceptions [33]. References to individual phonemes are only used in some examples and are not essential for following the discussion.

2.3 Hearing and psychoacoustic scales

The human speech production and hearing mechanisms are very well matched with each other and may have evolved in parallel [27]. The ear is especially responsive to those frequencies in the speech signal that contain the information most relevant to communication, i.e., in the 200-5600 Hz range. The listener can discriminate small differences in time and frequency found in speech sounds in this frequency range. In terms of the sound pressure level (SPL), the hearing threshold rises sharply with decreasing frequency below 1 kHz and with increasing frequency above 5 kHz. When measured 1 m from the lips, normal speech has an average SPL of roughly 60 dB. In the frequencies of greatest interest, the SPL is clearly above the normal threshold of hearing and well below the threshold of feeling [27].

The human ear can be divided into three main parts: *the outer ear, the middle ear, and the inner ear*. Figures 2.3 and 2.4 show a drawing and a simplified diagram, respectively, of the peripheral auditory system.

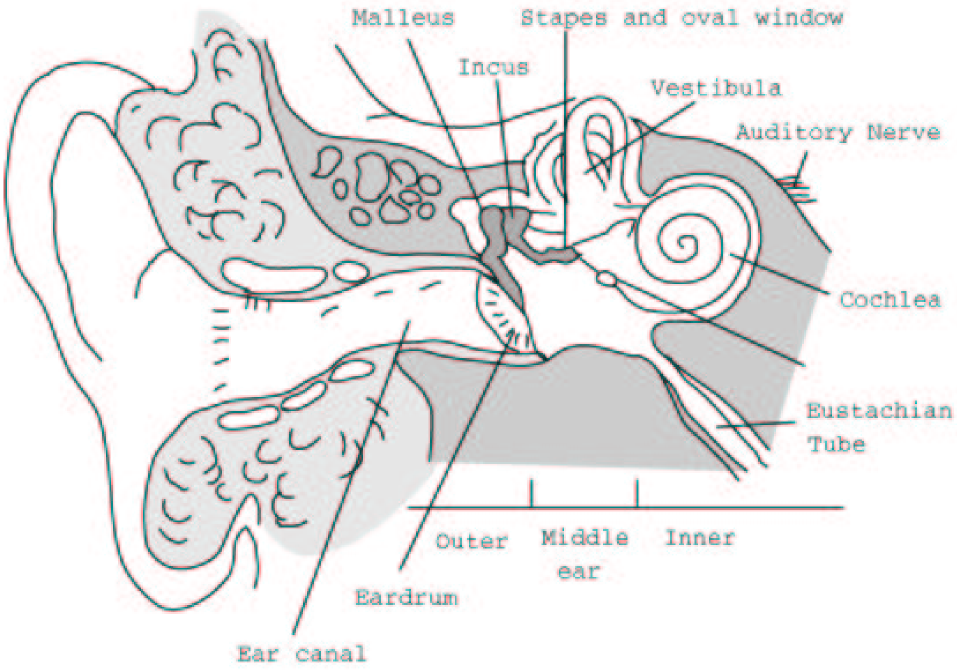

Figure 2.3: The structure of the peripheral auditory system (from [18]).

The outer ear consists of the *pinna* and the *ear canal* and is separated from the middle ear by the *eardrum*. The function of the outer ear is simply to act as an acoustic filter passing the sound waves into the middle ear. The pinna is relevant for localization of sound; by its asymmetric shape, it makes the ear more sensitive to sounds coming from in front of the listener than to those coming from behind and thus causes these directions to be perceived

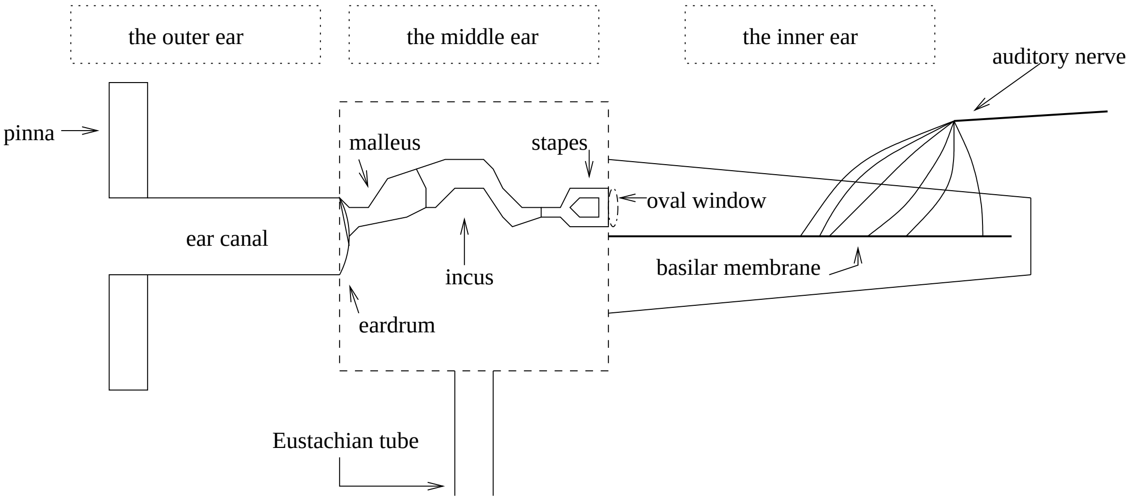

Figure 2.4: A simplified diagram of the human ear (after [16]). The cochlear part has been drawn straight instead of coiled.

differently. The ear canal can be approximated as a simple acoustic tube open at one end (pinna) and closed at the other (eardrum). The canal in an adult is between 2 and 3 cm in length and about 0.7 cm in width. Acting as a quarter-wavelength resonator, it amplifies energy in the 3-5 kHz range.

The eardrum marks the beginning of the middle ear, an air-filled cavity of about 6 cm^3 that contains the *ossicular bones*: *malleus* (hammer), *incus* (anvil), and *stapes* (stirrup). Their function in the hearing mechanism is to linearly transmit eardrum vibrations to the *oval window* membrane of the inner ear while also doing an acoustic impedance transformation. The liquid medium in the inner ear has about 4000 times higher an acoustic impedance than the air medium in the outer ear. The transformation is based primarily on the large area difference between the eardrum (about 65 mm^2) and the stapes (about 3 mm^2 [16]), but secondarily also on the lever action of the ossicular bones.

The *cochlea*, located in the inner ear, is a coiled tube filled with a gelatinous fluid called *endolymph*. Sound waves enter the cochlea from the middle ear via the oval window. Within the cochlea, the vibrations in the fluid cause the *basilar membrane* to start vibrating as well. On the basilar membrane lies the *organ of Corti* which contains about 30000 sensory *hair cells*, arranged in several rows along the length of the cochlea and the basilar membrane. Without going into too much detail, basilar membrane vibration in some location causes the affected hair cells to send electrical impulses up the neural fibers of the *auditory nerve*. The basilar membrane varies gradually in width and density along its length. In the beginning (near the oval window and the *round window*) it is narrow and stiff but near the apex it is compliant and massive. Each location on the basilar membrane responds differently to sounds in different frequencies. High frequency traveling waves in the inner ear resonate near the beginning of the basilar membrane, while low frequencies travel across the membrane and resonate in the apex end. In terms of the resonance location on the membrane, the frequency resolution of the basilar membrane is best at low frequencies.

A classical and fundamental concept in the study of hearing and sound perception is that of a *critical band* related to the frequency resolution of hearing. A critical band defines a frequency range in psychoacoustic experiments for which perception abruptly changes as a narrowband sound stimulus is modified to have frequence components beyond the band. When two competing sound signals pass energy through a critical-band filter, the sound with higher energy within the critical band dominates the perception and masks the other sound. Critical bandwidths can be measured by various slightly different psychoacoustic tests [16] [27]. Below 500 Hz, critical bandwidth is roughly constant at about 100 Hz. For higher frequencies, it increases proportional to the center frequency (roughly logarithmically above 1 kHz) reaching bandwidths of 700 Hz when the center frequency is near 4 kHz. An analytic expression for mapping the frequency f into critical-band rate z, or the *Bark scale*, is [27]

$$z = 13\tan^{-1}(0.76f/\mathrm{kHz}) + 3.5\tan^{-1}(f/(7.5\mathrm{kHz}))^2 \tag{2.1}$$

This mapping is shown graphically in figure 2.5. The frequency resolution is obviously nonuniform; dz/df is much greater, meaning better frequency resolution, at low frequencies than at higher frequencies. Recalling what was said about formants $F1$-$F3$ in section 2.2, it is readily seen that these perceptually most important formants are located exactly in the frequency range in which the frequency resolution of hearing is at its best.

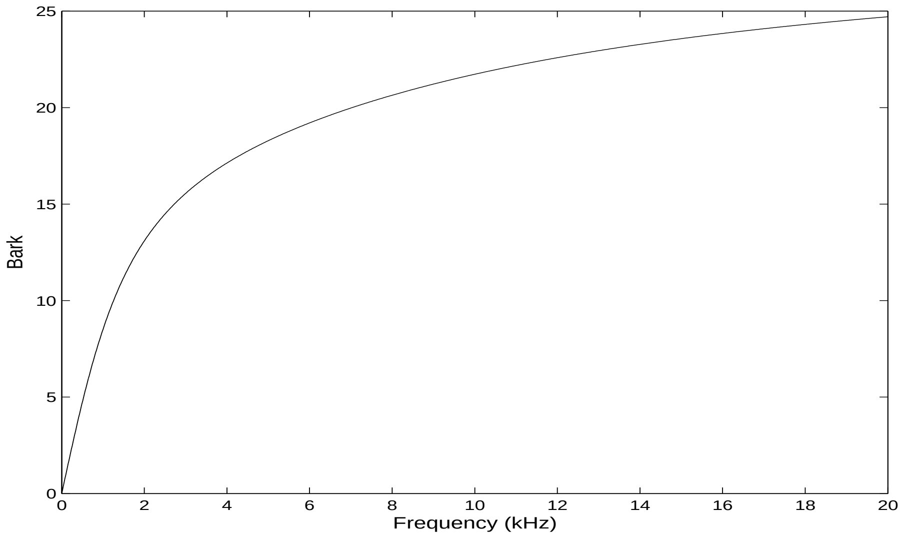

Figure 2.5: The Bark scale versus the frequency scale.

Chapter 3

Parametric speech spectrum modeling

3.1 Linear prediction

3.1.1 ARMA processes and pole-zero models

In time series analysis, the basic building block is a *white noise process*. Such a process $\{\varepsilon_n\}_{n=-\infty}^{\infty}$ satisfies

$$
\begin{aligned}
E(\varepsilon_n) &= 0 && (3.1)\\
E(\varepsilon_n^2) &= \sigma^2 && (3.2)\\
E(\varepsilon_n \varepsilon_{n+k}) &= 0 \qquad \text{for } k \neq 0 && (3.3)
\end{aligned}
$$

That is, the process has mean zero and variance σ^2 and the samples are uncorrelated across time. If

$$\varepsilon_n \sim N(0, \sigma^2) \tag{3.4}$$

also holds, it is a *Gaussian white noise process*. These specifications are used in defining three types of processes.

In a *q-th order moving average process*, denoted MA(q), each new value is a linear combination of a mean term μ and q past values of the noise process:

$$s_n = \mu + \varepsilon_n + b_1\varepsilon_{n-1} + b_2\varepsilon_{n-2} + \ldots + b_q\varepsilon_{n-q} \tag{3.5}$$

.

In a *p-th order autoregressive process*, denoted AR(p), each new value is a linear combination of a constant term c, p past values of the process itself, and a noise term ε_n:

$$s_n = c + a_1 s_{n-1} + a_2 s_{n-2} + \ldots + a_p s_{n-p} + \varepsilon_n \tag{3.6}$$

An ARMA(p,q) process is a combination of AR and MA processes and includes both the autoregressive and the moving average terms:

$$s_n = c + a_1 s_{n-1} + \ldots + a_p s_{n-p} + \varepsilon_n + b_1 \varepsilon_{n-1} + \ldots + b_q \varepsilon_{n-q} \tag{3.7}$$

From equation (3.7) it is easy to see that both AR(p) and MA(q) processes can be obtained as special cases of an ARMA(p,q) process by setting either $b_l = 0$, $l = 1, \ldots, q$ or $a_k = 0$, $k = 1, \ldots, p$, respectively.

These processes can be used for modeling wide-sense-stationary signals. In speech signal processing, only air pressure variations around the average atmospheric pressure are measured and analyzed; the atmospheric pressure does not have any meaning from the communication point of view and does not appear in the recorded signal. Therefore the long term average of the signal is zero and it is natural to set the constant term to zero, $c = 0$, in equation (3.7). The innovation sequence ε_n will also be replaced with a more general input signal Gu_n, where u_n is an unknown input and G denotes the gain factor. With these modifications, (3.7) can be written

$$s_n = \sum_{k=1}^{p} a_k s_{n-k} + Gu_n + \sum_{l=1}^{q} b_l G u_{n-l} \tag{3.8}$$

A z transform on both sides of (3.8) gives the transfer function of the system generating the signal as

$$H(z) = \frac{S(z)}{U(z)} = G \frac{1 + \sum_{l=1}^{q} b_l z^{-l}}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{3.9}$$

where $S(z)$ and $U(z)$ are the z transforms of the s_n and u_n, respectively.

The ARMA model of equation (3.9) corresponds to a *pole-zero signal model*. The roots of the numerator polynomial are the zeros and the roots of the denominator polynomial are the poles. When $H(z)$ contains only zeros or only poles the model is known as an all-pole or an all-zero model, respectively. Thus, an AR model is an all-pole model and a MA model is an all-zero model. Digital filters that realize an all-zero model are FIR filters; filters that realize an all-pole or a pole-zero model are IIR filters.

3.1.2 The all-pole model

AR, or all-pole models are particularly well suited for modeling speech signals, because they are easy to estimate and have a connection to the speech production mechanism. An all-pole model can be viewed as a digital filter representation of a lossless acoustic tube model, that is a simplified physical model of the speech production system [24]. The all-pole model for speech production is expressed in the frequency domain by the system function

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} = \frac{G}{A(z)} \tag{3.10}$$

which specifies the *synthesis filter*. This is the IIR filter assumed to have generated the observed speech signal. The polynomial $A(z)$ in the denominator of (3.10) is the system function for a FIR filter known as the *inverse filter*. The term G in the numerator is the input gain of the synthesis filter model.

$H(z)$ can be equivalently expressed in the time domain by the difference equation specifying an $AR(p)$ process:

$$s_n = \sum_{k=1}^{p} a_k s_{n-k} + G u_n \tag{3.11}$$

The autoregressive parameters a_k, $k = 1, \dots, p$, are also known as *predictor coefficients*; when p latest signal samples are known, the model of (3.11) can be used to form a prediction $\hat{s}_n$ of the next sample:

$$\hat{s}_n = \sum_{k=1}^{p} a_k s_{n-k} \tag{3.12}$$

This explains why all-pole/AR modeling is also known as *linear prediction* (*LP*) or *linear predictive coding* (*LPC*). The number of predictor coefficients p is the *prediction order*.

The prediction error or *residual* e_n is the difference between the actually observed speech samples s_n and the model-based predictions $\hat{s}_n$, according to (3.12):

$$e_n = s_n - \hat{s}_n = s_n - \sum_{k=1}^{p} a_k s_{n-k} \tag{3.13}$$

From 3.13 it can be seen that e_n is the signal obtained by filtering s_n with the inverse filter $A(z)$, also known as the prediction error filter. Comparing equations (3.11) and (3.13) shows that if the model is correct, then $e_n = G u_n$.

In the following, speech signal is treated as deterministic and the method of least squares is applied for solving the model parameters a_k. On the basis of equation (3.13), the total squared prediction error E (with a predictor of order p) can be calculated from

$$E = \sum_n e_n^2 = \sum_n (s_n - \sum_{k=1}^{p} a_k s_{n-k})^2 \tag{3.14}$$

which is minimized with respect to the model parameters by setting

$$\frac{\partial E}{\partial a_i} = 0, \qquad 1 \leq i \leq p \tag{3.15}$$

Expanding the square in (3.14) and setting the partial derivatives to zero as in (3.15) leads to a set of equations known as the *normal equations* [22]:

$$\sum_{k=1}^{p} a_k \sum_n s_{n-k} s_{n-i} = \sum_n s_n s_{n-i}, \qquad 1 \leq i \leq p \tag{3.16}$$

These p equations in p unknowns can be solved to obtain the set of predictor coefficients $\{a_k\}$, $1 \leq k \leq p$, that minimize E. The minimum value for the total squared prediction error can be shown to be

$$E_p = \sum_n s_n^2 - \sum_{k=1}^{p} a_k \sum_n s_n s_{n-k} \tag{3.17}$$

The range of summation over n in equations (3.14)-(3.17) was left unspecified. The choice of the range of summation leads to the two basic methods of linear prediction: the *autocorrelation method* and the *covariance method*. These will be discussed next.

3.1.3 Block estimation

Autocorrelation method

In the autocorrelation method, the error in (3.14) is minimized over an infinite interval, $-\infty \leq n \leq \infty$. In practice, of course, it is not possible to process an infinite signal. Furthermore, one is probably only interested in some small interval, say $0 \leq n \leq N-1$, of the signal at a time. Let x_n be the full-length signal of which we only want to process a small portion s_n starting from the time index m. Let w_n be a length N window function which is nonzero only for $0 \leq n \leq N-1$. A Hamming window is almost always used for w_n, in this thesis also. We obtain s_n by applying the window function on x_n:

$$s_n = \begin{cases} x_{m+n} w_n & 0 \leq n \leq N-1 \\ 0 & \text{otherwise} \end{cases} \tag{3.18}$$

Since s_n is zero outside the analysis frame, equations (3.16) and (3.17) reduce to

$$\sum_{k=1}^{p} a_k R(i-k) = R(i), \qquad 1 \leq i \leq p \tag{3.19}$$

and

$$E_p = R(0) - \sum_{k=1}^{p} a_k R(k) \tag{3.20}$$

where

$$R(i) = \sum_{n=0}^{N-1} s_n s_{n-i} = \sum_{n=i}^{N-1} s_n s_{n-i} \tag{3.21}$$

is the short term autocorrelation function for the windowed signal. Taking into account the fact that $R(i)$ is an even function, $R(i) = R(-i)$, equations (3.19) can be expressed in matrix form as

$$\begin{pmatrix} R(0) & R(1) & R(2) & \dots & R(p-1) \\ R(1) & R(0) & R(1) & \dots & R(p-2) \\ R(2) & R(1) & R(0) & \dots & R(p-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R(p-1) & R(p-2) & R(p-3) & \dots & R(0) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} R(1) \\ R(2) \\ R(3) \\ \vdots \\ R(p) \end{pmatrix}$$

or in a shorter form as

$$\boldsymbol{Ra} = \boldsymbol{r} \tag{3.22}$$

The autocorrelation matrix $\boldsymbol{R}$ is a *Toeplitz* matrix: all elements along any diagonal are equal. It is also a symmetric matrix. Moreover, the right hand side elements are the same as those found in the matrix on the left hand side. All this special structure allows for an efficient solution algorithm known as *Levinson-Durbin recursion* [29]. When processing a block of N samples, the autocorrelation method using Levinson-Durbin recursion requires about pN multiplications for computing the autocorrelation matrix and about p^2 operations for solving the matrix equations [29]. The total computational cost is thus on the order of $pN + p^2$ [23]. Of course, the computational cost for block estimation methods increases if the signal is processed in successive overlapping blocks as is usually done in practice.

The Levinson-Durbin recursion also leads to another formulation of the linear prediction problem, namely the *lattice* methods. A popular example of lattice methods based on block processing is *Burg's method* [23]. Its efficiency is comparable to the covariance method (to be discussed next).

Covariance method

In the covariance method, the error in (3.14) is minimized only over a finite interval, $p \leq n \leq N-1$. It is assumed that the signal is available over the interval $0 \leq n \leq N-1$. Equations (3.16) and (3.17) reduce to

$$\sum_{k=1}^{p} a_k \varphi_{ik} = \varphi_{i0}, \qquad 1 \leq i \leq p \tag{3.23}$$

and

$$E_p = \varphi_{00} + \sum_{k=1}^{p} a_k \varphi_{0k} \tag{3.24}$$

where

$$\varphi_{ij} = \sum_{n=p}^{N-1} s_{n-i} s_{n-j} \tag{3.25}$$

are correlation-like values calculated from the data within the signal frame. In matrix form, equations (3.23) can be written

$$\begin{pmatrix} \varphi_{11} & \varphi_{12} & \varphi_{13} & \cdots & \varphi_{1p} \\ \varphi_{21} & \varphi_{22} & \varphi_{23} & \cdots & \varphi_{2p} \\ \varphi_{31} & \varphi_{32} & \varphi_{33} & \cdots & \varphi_{3p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \varphi_{p1} & \varphi_{p2} & \varphi_{p3} & \cdots & \varphi_{pp} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \cdots \\ a_p \end{pmatrix} = \begin{pmatrix} \varphi_{10} \\ \varphi_{20} \\ \varphi_{30} \\ \cdots \\ \varphi_{p0} \end{pmatrix}$$

or more concisely as

$$\boldsymbol{Ca} = \boldsymbol{\varphi} \tag{3.26}$$

Also the covariance method matrix is symmetric, since $\varphi_{ij} = \varphi_{ji}$, but it is not Toeplitz. Also, the right-hand-side terms do not appear on the left hand side. The covariance method equations are usually solved using decomposition methods such as the Cholesky decomposition. This takes on the order of $(1/6)p^3 + (3/2)p^2$ operations [29]. The computation of the correlation matrix requires about pN multiplications as with the autocorrelation method. When processing a single block of N samples, the computational load caused by the covariance method is on the order of $pN + (1/6)p^3 + (3/2)p^2$ [23]. It is thus computationally somewhat less efficient than the autocorrelation method.

3.1.4 Adaptive estimation

Block estimation methods try to solve the coefficient vector $\boldsymbol{a}$ (equations (3.22) and (3.26)) directly by assuming the signal stationary within the estimation frame. The coefficient vector can also be estimated in an adaptive fashion. There are many adaptive estimation techniques, but in the present study we shall be concerned only with FIR adaptive filters based on *gradient descent* principles. They are computationally efficient, easily controllable in terms of stability, simple to implement, well understood and discussed in the literature, and usually perform adequately. Tuning the adaptation rate of these methods is a compromise between the ability to track rapid changes in signal statistics and the amount of *misadjustment*, which is the excessive mean square error occurring even after the filter has converged on a stationary signal [8].

Least-Mean-Square (LMS) algorithm

A direct form (transversal) FIR filter structure has the following type of *update equation*:

$$\boldsymbol{a}_{n+1} = \boldsymbol{a}_n + \Delta\boldsymbol{a}_n \tag{3.27}$$

where $\Delta\boldsymbol{a}_n$ is a correction applied to the filter coefficients $\boldsymbol{a}_n$ at time n to form the new set of coefficients for the next sample, $\boldsymbol{a}_{n+1}$. The sequence $\{\boldsymbol{a}_n\}$ can be sampled at the desired times n to obtain estimates for the coefficient vector $\boldsymbol{a}$.

One FIR adaptive filtering method is based on the popular *least-mean-square (LMS)* adaptation algorithm, also known as the Widrow-Hoff algorithm after those who first introduced it in [35]. The LMS approach to linear prediction uses a direct-form filter structure, whose tap weights are adapted continuously with each incoming sample. The LMS algorithm itself is more general and can also be used in other applications than adaptive signal filtering, such as designing linear pattern classifiers based on a minimum-squared error criterion [3]. The general LMS algorithm is as follows.

1. At each time n, the coefficients $\boldsymbol{a}_n$ are used to form a prediction of the *desired response* d_n based on the value of the vector process $\boldsymbol{x}_n$:

$$\hat{d}_n = \boldsymbol{a}_n^T \boldsymbol{x}_n$$

2. The prediction error is computed:

$$e_n = d_n - \hat{d}_n$$

3. The coefficients are updated based on the prediction error e_n:

$$\boldsymbol{a}_{n+1} = \boldsymbol{a}_n + \mu e_n \boldsymbol{x}_n$$

In the case of linear prediction, the desired response is the signal being modeled, $d_n = s_n$. In ordinary LP, at time n, the vector $\boldsymbol{x}_n$ contains p past values of the signal: $\boldsymbol{x}_n = (s_{n-1}, s_{n-2}, \ldots, s_{n-p})^T$.

It can be shown that the LMS algorithm converges in the mean square for stationary processes if the step size parameter μ satisfies $0 < \mu < \mathrm{tr}(\boldsymbol{R}_d)$, where the upper bound is the trace of the autocorrelation matrix of the signal d_n [7]. Here, we use a variant known as *normalized LMS* [7]; in this case the update equation in step 3 becomes

$$\boldsymbol{a}_{n+1} = \boldsymbol{a}_n + \beta \frac{\boldsymbol{x}_n}{\epsilon + \|\boldsymbol{x}_n\|^2} e_n$$

where β is the normalized step size with $0 < \beta < 2$ and ϵ is a small positive number to prevent the amount of update from assuming too large values when $\boldsymbol{x}_n$ becomes small. It can be shown that the normalized LMS algorithm converges in the mean square if $0 < \beta < 2$ [7]. The normalization diminishes *gradient noise amplification* occurring with basic LMS when the norm of $\boldsymbol{x}_n$ becomes large [7]. In addition, it makes the selection of the step size parameter a little easier.

Let it be noted that LMS is known to generally not perform well with signals that have a large *eigenvalue spread* (the ratio of the largest eigenvalue to the smallest eigenvalue of the signal autocorrelation matrix) such as speech; transform domain modifications for alleviating this problem are suggested in [26]. Such techniques are not used here, but in the experimental part of this thesis it is checked if the situation is any better when the tap inputs have been transformed by an *allpass transform* (section 3.2).

LMS is quite efficient computationally; the number of operations requires per sample is proportional to p and each sample is processed once. The total computational load for processing a sequence of N samples is thus on the order of pN and the cost does not increase no matter how often the coefficient estimates are sampled.

For more details on LMS and its convergence, see [7] [8] [36].

Gradient Adaptive Lattice (GAL)

The *gradient adaptive lattice (GAL)* algorithm was first introduced in [5]. It is a FIR adaptive filter that uses a lattice structure instead of a direct form structure. Due to the fact that the lattice filter orthogonalizes the input process, the convergence of the gradient adaptive lattice filter is typically faster than that of LMS [7]. The gradient adaptive lattice requires approximately twice as many operations per update than required for LMS. However, like in LMS, the computational load does not increase if the vectors are sampled more frequently and the total computational cost is still small in comparison with e.g. the autocorrelation method with considerably overlapping model estimation frames. For details on the gradient adaptive lattice method, see [7] [8].

3.2 Warped linear prediction (WLP)

3.2.1 Frequency warping by an allpass transform

Frequency warping, as it is perceived, is a transformation from a linear, uniform resolution frequency scale (Hz) to a nonuniform resolution frequency scale. Frequency warping is mostly applied to signal models and spectral representations. A straightforward way to implement the warping is to replace the unit delay elements in a DSP structure with first order allpass filters, having transfer functions

$$D(z) = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}}, \tag{3.28}$$

where $-1 < \lambda < 1$ is the warping coefficient (amount of warping). When the allpass sections $D(z)$ are arranged in a chain, the allpass chain forms a dispersive delay line with frequency dependent delays. When λ is reduced to zero, allpass sections are transformed back to unit delays and no warping occurs.

When the warped signal is observed over the tap outputs of the allpass chain, its effective time window proves to be frequency dependent. Thus, the Fourier transform of the tap outputs results in nonuniform frequency representation. Similarly, the autocorrelation over the tap outputs leads to a warped autocorrelation function and can further be used for block-based linear predictive spectral estimation, in which case the resulting spectral envelope model is warped in frequency. By a proper choice of the warping parameter λ, the frequency warping can be made “nearly identical” to the psychoacoustic based Bark scale, thus optimizing the frequency resolution from the point of view of auditory perception [30]. This thesis is concerned with frequency-warped linear prediction whose frequency scale is an approximation of the Bark scale.

The phase response of the allpass filter is [12]

$$\tilde{\omega} = \omega + 2tan^{-1}\left(\frac{\lambda \sin \omega}{1 - \lambda \cos \omega}\right) \tag{3.29}$$

There is a so-called *turning point frequency* f_{tp}, which is the frequency at which the frequency resolution is unaffected by warping. Below f_{tp} the frequency resolution is higher and above f_{tp} the frequency resolution is lower than in a normal system having uniform frequency resolution. The turning point frequency in Hz is given by [12]

$$f_{tp} = \frac{f_s}{2\pi} cos^{-1}(\lambda) \tag{3.30}$$

where f_s is the sampling frequency in Hz.

A formula for the coefficient λ in the allpass transform that gives a good approximation for Bark frequency scale has been derived by Smith and Abel [31] (with a typographical error in the formula in the article corrected in [12]):

$$\lambda_{f_s} \approx 1.0674(\frac{2}{\pi} tan^{-1}(0.06583 f_s/1000))^{1/2} - 0.1916 \tag{3.31}$$

where f_s is the sampling rate in Hz.

3.2.2 Conversion of conventional LP to WLP

A more general point of view is appropriate for the discussion of LP and WLP in the rest of the thesis. Härmä [10] has discussed linear prediction with modified filter structures. In this case the associated inverse and synthesis filters consist of branches with generalized filter elements $D_k(z)$. The transfer function of the model is

$$H(z) = \frac{G}{A(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k D_k(z)} \tag{3.32}$$

The generalized inverse filter of this model is shown in figure 3.1. The network for solving the coefficients of the generalized model is shown in figure 3.2.

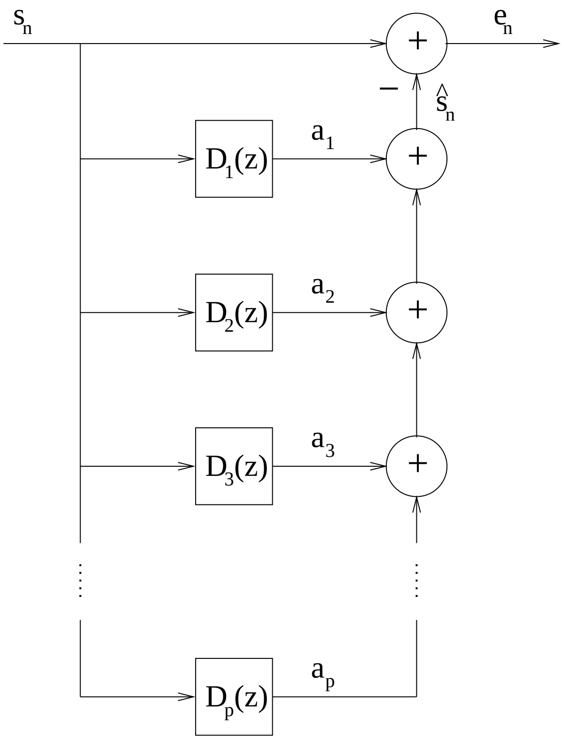

Figure 3.1: The generalized LP inverse filter (after [10]).

Denote the output of the kth filter element at time instant n by $y_{k,n}$, so that

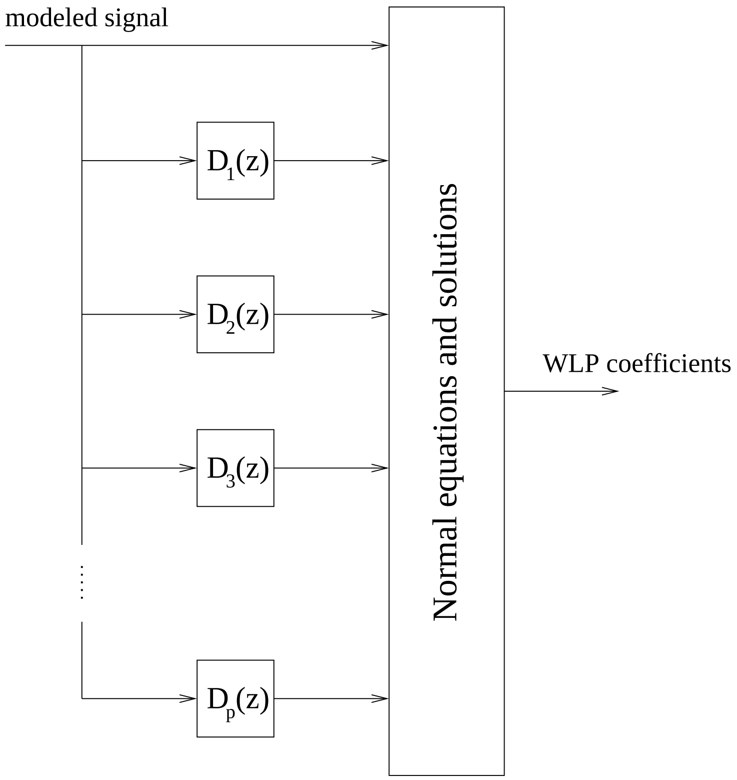

Figure 3.2: The generalized LP network for computing the model coefficients (after [10]).

$$y_{k,n} = \begin{cases} s_n & k = 0 \\ \sum_{m=0}^{\infty} d_{k,m} s_{n-m} & 1 \leq k \leq p \end{cases} \tag{3.33}$$

where $d_{k,m}$ is the mth sample of the (potentially infinite) impulse response of $D_k(z)$. Then, (3.32) can equivalently be written in the time domain as

$$s_n = \sum_{k=1}^{p} a_k y_{k,n} + G u_n \tag{3.34}$$

A particularly useful situation occurs if each element $D_k(z)$, $1 \leq k \leq p$, is a cascade of k identical filters. In this thesis, the filters $D_k(z)$ will always be assumed to be cascades of *first-order allpass* filter sections (transfer function given by equation (3.28)) such that

$$D_k(z) = \prod_{i=1}^{k} \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}} \tag{3.35}$$

Because there is no feedback between individual allpass elements, the filter outputs correspond to the outputs of the different stages of an allpass chain. Using the allpass transfer function (3.28), the cascade analysis model in figure 3.3 represents the allpass chain model.

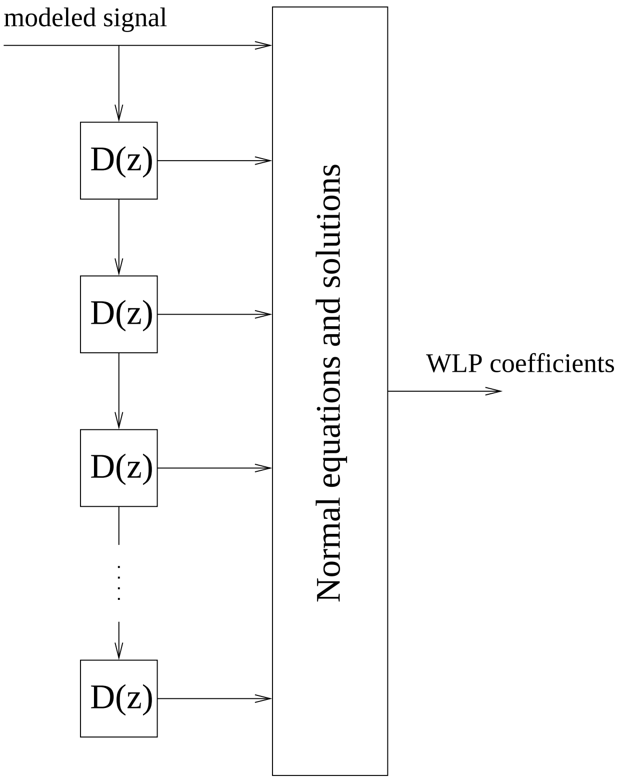

Figure 3.3: The cascade network for computing the LP model coefficients.

The individual impulse responses $d_{k,m}$ are dependent on both the number of allpasses in cascade before the kth output as well as the value of the warping coefficient λ.

The process of (3.34) is no longer necessarily an $AR(p)$ process, but instead potentially an $AR(\infty)$ process. It is however obvious that each speech sample s_n can be related by linear regression to the $(p \times 1)$ vector containing the state of the allpass chain, that is, the instantaneous outputs $(y_{1,n}, y_{2,n}, \ldots y_{p,n})^T$. Thus, a similar least squares minimization of the total squared residual

$$E = \sum_{n=0}^{N-1} e_n^2 = \sum_{n=0}^{N-1} (s_n - \sum_{k=1}^{p} a_k y_{k,n})^2 \tag{3.36}$$

as in section 3.1 leads to a set of normal equations given by

$$\sum_{k=1}^{p} a_k \sum_n y_{k,n} y_{i,n} = \sum_n s_n y_{i,n}, \qquad 1 \leq i \leq p \tag{3.37}$$

and the minimum total squared error is given by

$$E_p = \sum_n s_n^2 - \sum_{k=1}^{p} a_k \sum_n s_n y_{k,n} \tag{3.38}$$

Let it also be noted that substitution of (3.33) in equations (3.37)-(3.38) and changing the orders of summation leads to

$$\sum_{k=1}^{p} a_k \sum_{m=0}^{\infty} \sum_{q=0}^{\infty} d_{k,m} d_{i,q} \sum_n s_{n-m} s_{n-q} = \sum_{m=0}^{\infty} d_{i,m} \sum_n s_n s_{n-m}, \qquad 1 \leq i \leq p \tag{3.39}$$

and

$$E_p = \sum_n s_n^2 - \sum_{k=1}^{p} a_k \sum_{m=0}^{\infty} d_{k,m} \sum_n s_n s_{n-m} \tag{3.40}$$

respectively.

The special case of ordinary LP is achieved by making each $D(z)$ in equation (3.32) and figure 3.3 correspond to a unit delay element, $D(z) = z^{-1}$. From the allpass transfer function (3.28), it can be seen that this is achieved simply by setting $\lambda = 0$. In this case, each impulse response $d_{k,n}$ is nonzero for only a single sample and thus finite:

$$d_{k,n} = \begin{cases} 1 & n = k \\ 0 & n \neq k \end{cases} \tag{3.41}$$

From (3.41) and (3.33) it follows that in the non-warped case, $y_{k,n} = s_{n-k}$. Thus, equations (3.32) and (3.34) are reduced to equations (3.10) and (3.11). Equations (3.39) and (3.40) are similarly reduced to equations (3.16) and (3.17).

Allpass transforms were discussed in section 3.2.1. By setting $|\lambda| > 0$ in (3.28) and (3.35), a nonlinear frequency mapping occurs. This leads to *warped linear prediction (WLP)*. In the WLP model, the linear predictor generating the observed samples operates not on a vector of p past signal values but instead on a sample from another p-dimensional vector process, obtained from the signal by an allpass transform which effectively warps the frequency axis.

It is shown in [10] that in the case of *warped autocorrelation method*, when the residual minimization is done over an infinite interval, equations (3.37)-(3.38) can be simplified to

$$\sum_{k=1}^{p} a_k C(i-k) = C(i), \qquad 1 \leq i \leq p \tag{3.42}$$

and

$$E_p = C(0) - \sum_{k=1}^{p} a_k C(k) \tag{3.43}$$

where

$$C(i) = \sum_{n=0}^{N-1} s_n y_{i,n} \tag{3.44}$$

where s_n is the suitably windowed signal. Note the similarity with (3.19)-(3.21) but note also the difference in the limits of summation in (3.21) and (3.44). This is because unlike s_{n-i}, $y_{i,n}$ is not guaranteed to be zero outside the interval $i \leq n \leq N-1$.

Substituting (3.33) in (3.44) we see that

$$C(i) = \sum_{m=0}^{\infty} d_{i,m} \sum_{n=0}^{N-1} s_n s_{n-m} = \sum_{m=0}^{\infty} d_{i,m} R(m) \qquad 0 \leq i \leq p \tag{3.45}$$

Thus, the signal can also be warped in the *autocorrelation domain* [32] [21]. In practice, the warped correlation terms $C(k), k = 0 \ldots p$ can be obtained by multiplying the length L autocorrelation sequence $R(k), k = 1 \ldots L-1$ by a $(p+1 \times L)$ matrix having as ith row the impulse response of a cascade of $i+1$ first-order allpasses. In this form, warping in the autocorrelation domain requires a large value of L in order to give good results. It is not discussed further in this thesis.

The *warped covariance method* equations are given by

$$\sum_{k=1}^{p} a_k C(i,k) = C(i,0), \qquad 1 \leq i \leq p \tag{3.46}$$

and

$$E_p = C(0,0) - \sum_{k=1}^{p} a_k C(0,k) \tag{3.47}$$

where

$$C(i,j) = \sum_{n=p}^{N-1} y_{i,n} y_{j,n} \tag{3.48}$$

Note again the similarity with (3.23)-(3.25). Actually, there are several possibilities for the choice of the limits of summation in this case. The definition of equation (3.48) will be used in this thesis. The choice is rather arbitrary, because according to the present formulation, $y_{i,n}$ can in theory be nonzero for any $n \geq 0$ and for all $i \geq 1$.

In summary, the WLP synthesis model can be described as an autoregressive model with unit delay elements replaced by first-order allpass filters. It was seen that both the autocorrelation method and covariance method of conventional linear prediction could be warped by making this modification to the analysis structures. LMS for WLP can likewise be implemented by replacing the unit delays in the direct form filter (tapped delay line) by allpasses. Taking a slightly different viewpoint, the WLP synthesis model is viewed as a linear regression model of the signal samples on vectors containing the states of the allpass chain, $\boldsymbol{x}_n = (y_{1,n}, y_{2,n}, \ldots y_{p,n})^T$, and the general formulation from section 3.1.4 is used to find the coefficients. The warped version of the GAL method is obtained similarly and has been introduced in [13]. There are, however, certain special techniques for which the WLP implementation is not as straightforward [12].

The experimental part of this thesis studies the warped versions of the autocorrelation method, the covariance method, LMS, and GAL.

Linear prediction on a warped frequency scale using the allpass transform was first discussed systematically by Strube [32]. The technique may prove useful in any digital signal processing application where emphasizing the spectrum at some frequency range is desirable. Perhaps the most relevant frequency scale in speech and audio applications is the Bark scale. WLP with Bark mapping is a simple way to incorporate a rudimentary form of auditory modeling in the LP framework. Applications include wideband audio coding [20] [11] [12] and speech synthesis [19]. Karjalainen [17] has pointed out some system-level similarities with Bark-WLP and the peripheral auditory system.

3.3 Other auditory speech analysis methods

Models that in some way or another characterize the auditory spectrum are widely used as analysis methods in automatic speech recognition. This kind of feature generation is intuitively appealing, as it tries to imitate the way humans recognize speech. It has been found to improve the recognition accuracy compared to normal spectrum modeling. The most popular analysis method for speech recognition is based on the *cepstrum* [2], but with a nonlinear frequency axis following the Bark of mel scale. The mel-frequency cepstral coefficients (MFCC) provide a representation for speech spectra which incorporates some aspects of audition. A common approach simulates critical-band filtering with a set of 20 triangular windows, with different widths and center frequencies, used on the logarithmic spectrum [27].

The cepstrum is often used as the final feature representation form for LP or WLP models when they are used in speech recognition front-ends. The LP/WLP models may be converted into cepstral representations by simple formulas [24]. WLP-derived cepstra compared favorably against the MFCC, in terms of both the recognition accuracy and the computational requirements, in one limited vocabulary recognition experiment [34].

Hermansky [9] has discussed another approach to incorporate the perceptual aspects in classical LP analysis. Perceptual linear prediction (PLP) is a rather sophisticated method that uses not only the critical-band spectral resolution, but also the *equal-loudness curve* and the *intensity-loudness power law* in producing an approximation of the true auditory spectrum [9]. In comparison, WLP is a straightforward modification of LP that only models the critical-band spectral resolution. Both the MFCC and the PLP representations performed slightly better than WLP in one continuous speech recognition experiment, but the differences were not very large [33].

Chapter 4

Measures for evaluating spectrum models

The purpose in the experimental part of this work is to analyze the behavior of different WLP estimation techniques with varying inputs and analysis parameters. This is done by objectively evaluating the quality of the WLP models both in the time domain (prediction error) and in the frequency domain (modeling of the general shape of the spectrum). A total of five technical evaluation measures were chosen after careful consideration. They are

- Normalized residual energy, full band (up to the Nyquist frequency), V_N
- Normalized warped residual energy, full band (up to the Nyquist frequency), V_W
- Normalized residual energy at low frequencies (up to the turning point frequency), V_L
- Flatness of the residual spectrum, full band (up to the Nyquist frequency), F_N
- Flatness of the residual spectrum at low frequencies (up to the turning point frequency), F_L

All these measures can be used in both the normal and the frequency-warped case. The full-band measures V_N and F_N do not account for the frequency warping in any way. The V_W measure is also computed for the full band of the spectrum, but has been compensated for the theoretical gain factor in the warped inverse filter mentioned in [12]. The band-limited measures V_L and F_L use information only up to the turning point frequency of the allpass transform. It is just this band that is emphasized in WLP.

These measures are next introduced in detail. The notation related to LP/WLP modeling is the same as in chapter 3.

4.1 Normalized residual energy measures

4.1.1 The V_N measure

The normalized prediction error or normalized residual energy V_N is a traditional performance measure for linear prediction [1] [22] [29]. It is also known, usually in a slightly different logarithmic form, as the *prediction gain* [15].

First we assume that the *model evaluation* window is the same as that used for model estimation. For the unwarped autocorrelation method V_N is then defined as [1] [22]

$$V_N = \frac{E_p}{R(0)} = 1 - \frac{1}{R(0)} \sum_{k=1}^{p} a_k R(k) \tag{4.1}$$

and for the unwarped covariance method as [1]

$$V_N = \frac{E_p}{c_{00}} = 1 - \frac{1}{c_{00}} \sum_{k=1}^{p} a_k c_{0k} \tag{4.2}$$

Thus,V_N is the minimized residual energy (see section 3.1.2) normalized to the energy of the original signal frame. For the lattice methods, a comparable definition is obtained from the derivation of the Levinson-Durbin recursion as [22]

$$V_N = \prod_{i=1}^{p} (1 - k_i^2) \tag{4.3}$$

With the LP residual denoted by e_n, a general definition applying to all techniques is [29]

$$V_N = \frac{\sum_n e_n^2}{\sum_n s_n^2} \tag{4.4}$$

provided the limits of summation in both the numerator and the denominator are properly chosen. Let it be noted that in the ordinary autocorrelation method, the normalized error has special significance in the frequency domain because the *minimum* error equals the model gain squared, $E_p = G^2$. As a further consequence, it can be shown that in this case V_N can be interpreted as the *flatness* of the model spectrum [22].

The problem is to generalize the above definitions in order to deal with warped linear prediction. Consider equation (3.34) as a general multiple linear regression model with observations over some interval in n. All the summations over n in this section are done over this same interval. The residual from the regression is

$$e_n = s_n - \sum_{k=1}^{p} a_k y_{k,n} \tag{4.5}$$

and the sum of the squared residuals from the order p predictor is given by

$$E_p = \sum_n e_n^2 \tag{4.6}$$

According to multiple linear regression theory, the popular R^2 statistic for this model can be expressed as [25]

$$R^2 = \frac{S_s - E_p}{S_s} \tag{4.7}$$

where S_s is defined as

$$S_s = \sum_n s_n^2 \tag{4.8}$$

for *uncentered* R^2 and as

$$S_s = \sum_n s_n^2 - \frac{1}{N}(\sum_n s_n)^2 \tag{4.9}$$

for *centered* R^2 [6]. If s_n has a mean close to zero, for a large enough N it should not matter much which definition of S_s is used. The R^2 statistic answers to the question, what proportion of the variability in the observed random variable can be explained by the fitted regression model? Values of R^2 near 1 are generally taken as an indication that the model explains (or predicts) the data well.

Substituting the sum of squared residuals from (4.6) and the uncentered R^2 from (4.8) into (4.7), we obtain

$$V_N = 1 - R^2 = \frac{E_p}{S_s} = \frac{\sum_n e_n^2}{\sum_n s_n^2} \tag{4.10}$$

This is the same as the definition of V_N given by equation (4.4). The normalized error according to its conventional definition is thus directly related to the goodness of fit of a regression model to the data. It is therefore readily applicable to the frequency-warped case, since WLP is also linear regression modeling; however, this measure only measures the mathematical fit and thus does not reward the inherent frequency weighting in WLP.

Before the V_N measure can be used, the limits of summation in (4.4) (or (4.10)) need to be specified. Normally, these limits are the same as those used by the least squares block estimation and are undefined for adaptive estimation. In this thesis, however, we want to compare different block least squares methods and adaptive gradient methods using the same criteria. Hence, the summation limits for V_N are determined by the length of an independently chosen *evaluation frame* and do not depend on the estimation method in any way. The same approach is used for all measures; the evaluation frame is independent of the estimation frame (which, of course, does not exist for the adaptive methods).

4.1.2 The V_W measure

This can be thought of as a kind of a warped version of the V_N measure. In [12], the warped inverse filter is reported to have an additional gain factor $g \neq 1$ not present in the ordinary LP inverse filter. The existence of the additional gain can be justified by the fact that a frequency-domain integration of the log magnitude spectrum of the inverse filter results in zero with ordinary LP [24], but not with WLP unless either the filter coefficients or the filter output (the residual) are divided by a certain term given by [12]

$$g = 1 - \sum_{k=1}^{p} a_k(-\lambda)^k \tag{4.11}$$

In practice, the WLP residual e_n is divided by g before computing the V_W measure by equation (4.4). Observe that this correction does not affect ordinary LP; because if $\lambda = 0$, then $g = 1$ and $V_W = V_N$.

In this work, also the *spectral tilt* caused by the warping was corrected before the computation of V_W, even though it does not affect a full-band energy measure such as this one. Correction of the spectral tilt is more important in the computation of the remaining three measures.

4.1.3 The V_L measure

This is a band-limited version of the V_N measure, in which only information up to the turning point frequency (given by equation (3.30)) is used. In theory, this means that e_n and s_n in equation (4.4) are replaced by their low-pass-filtered versions. However, first we must correct for the spectral tilt mentioned earlier. The residual spectrum from WLP is tilted because of the allpass transform. To correct for this, the residual must be high-pass filtered by the filter [12]

$$H_{tc}(z) = \frac{1 - \lambda z^{-1}}{\sqrt{1 - \lambda^2}} \tag{4.12}$$

When computing this measure in practice, after filtering e_n by (4.12) in order to compensate for the spectral tilt, the power spectrum of the resulting signal is estimated by Fast Fourier Transform (FFT). The resulting spectrum, denoted by E_n^f, is then truncated at the turning point frequency. The FFT estimate of the power spectrum of s_n is also truncated at the turning point frequency. Then the band-limited energies of the residual and the original signal are computed from the truncated power spectra by summation, applying Parseval's theorem, and the ratio of the spectral energies is taken as the band-limited normalized prediction error V_L.

4.2 Residual spectral flatness measures

4.2.1 The F_N measure

The *spectral flatness* of a signal x_n, whose power spectrum estimate is given by the frequency signal X_n, is defined as [24]

$$\gamma = \frac{(\prod_{n=1}^{N} X_n)^{1/N}}{\frac{1}{N}\sum_{n=1}^{N} X_n} \tag{4.13}$$

Here we consider the *flatness of the WLP residual spectrum*, and we don't want the inherent tilt to show up in the measure, so we substitute $X_n = E_n^f$ and use (4.13) to compute F_N. The flatter the spectrum of the residual, the better the quality of the spectrum model. Linear prediction tries to *whiten* the residual spectrum by capturing the important main structure (e.g. formants) in the model and leaving the irrelevant details (e.g. pitch period harmonics) in the residual. Thus the residual spectral flatness measures could also be called residual whiteness measures. Warped linear prediction whitens the residual spectrum more at some frequency bands than at others.

4.2.2 The F_L measure

Analogous to the relation of V_N and V_L, this is a band-limited version of F_N. The flatness of the residual spectrum is only considered at low frequencies up to the turning point frequency. In practice, the tilt-compensated FFT power spectrum estimate E_n^f is truncated at the turning point frequency and the resulting spectrum is substituted for X_n in equation (4.13) in order to compute F_L.

Part II

Experiments

Chapter 5

Test setup

This chapter begins the experimental part of this thesis. The main goals of the experiments were:

1. To determine good analysis parameters for the methods. If the analysis parameters are improperly chosen, the results may be quite poor. WLP in general differs from ordinary LP in terms of assignment of the poles to signal frequencies. With the emphasis on the most relevant speech frequencies, systematic WLP model order determination can be used to make the models as good at these frequencies as ordinary LP. Another important point is adjusting the adaptation rate of the adaptive techniques to make their time resolution correspond to block estimation with a given block length.

2. To compare the optimized performance of each of the two less commonly used adaptive methods agains the “standard” block methods with different types of speech inputs.

The speech material used in the simulations was 10 utterances of Finnish sentences spoken by a male speaker. The material was recorded in an anechoic room with high quality equipment at a 22050 Hz sampling rate. The material was subsequently manually segmented into phoneme-size units and labeled by a trained phonetician. The phonemic transcription was further refined by assigning separate labels to voiced and unvoiced /h/ and to the occlusion and burst segments of stop consonants. This transcription was used for dividing the speech analysis frames into four nested categories for the purposes of this study: *all frames* means all signal frames including silence (all excitation types); *speech frames* means frames containing some speech sound (not silence); *voiced frames* means frames that contain sounds with voiced excitation; and *vowel frames* means frames that contain vowels. The *all frames* category is the most variable while *vowel frames* is the most stationary.

A speech analysis system was built for the purposes of this thesis using Matlab (version 6.0.0.88 (R12)) [14]. In short, the first phase of using the system involves running a script that computes various acoustic features from speech signals and saves the results and analysis parameters in a database. This phase does the actual LP/WLP computation and model evaluation.

The second phase consists of making queries specifying the following:

- identity of the feature (one of the measures from chapter 4)
- analysis parameter sets of interest; typically, one parameter is varied and the others are held fixed
- phonetic content (e.g. all frames, only non-silent frames, only frames containing voiced speech etc.)

The retrieved data is then summarized e.g. by averaging over all the retrieved different frames and visualized e.g. by plotting the averaged values of the feature against the varying analysis parameter. Also more elaborate analysis is possible. The third phase handles summarization and visualization of the results.

A variety of analysis parameters affects the computation in the first phase. Of these, the following are relevant in this thesis:

- Sampling rate f_s (16000 Hz or 22050 Hz)
- Warping coefficient λ (0 for ordinary LP, see Table 5.1 for Bark-WLP)
- Shift interval between adjacent speech frames (3 ms)
- Linear prediction method (one of autocorrelation, covariance, gradient adaptive lattice (GAL), or least-mean-squares (LMS))
- Linear prediction order p
- Length of the LP estimation frame
- Step size in LMS
- Memory coefficient in GAL
- Length of the evaluation frame

The choice of the sampling rate f_s and the warping coefficient λ are related. When Bark scale warping is desired, the optimal λ depends on f_s. The turning point frequency f_{tp}, as

Table 5.1: Sampling rate, warping coefficient, and turning point frequency.

Sampling rate f_s (Hz)	Warping coefficient λ	Turning point frequency f_{tp} (Hz)
16000	0.576	2437
22050	0.646	3048

defined in 3.30 and used in the computation of V_L and F_L, depends on both f_s and λ. Table 5.1 shows the combinations of f_s, λ, and f_{tp} used in this thesis.

Before any LP or WLP model estimation, the speech signal was *pre-emphasized* with the FIR highpass filter

$$H_p(z) = 1 - 0.95z^{-1} \tag{5.1}$$

Chapter 6

Data reduction

The short-time speech spectrum can generally be represented as having an average density of one formant per kiloHertz [29]. Linear prediction tries to model the spectrum using a given prediction order, which is equivalent to the number of poles in the transfer function. In ordinary linear prediction, a common rule is that the model should have two poles for each formant in the signal bandwidth plus a few additional poles. Thus, as a simple rule of thumb, the optimal number of poles is given by the sampling rate in kiloHertz plus a small integer, typically 2-4. The additional poles are needed to model the general spectral shape, mostly due to the spectrum of the glottal waveform and to lip radiation characteristics [27], as well as to compensate for zeros in the spectrum due to antiresonances occurring in nasalized and unvoiced sounds.

Warped linear prediction can be used to generate spectral models on an auditory frequency scale. An auditory frequency scale emphasizes those regions of the spectrum carrying the most relevant information content of speech, in particular the lowest formants. The low frequencies are emphasized and the high frequencies are de-emphasized, in terms of modeling accuracy. Because modeling is concentrated on the perceptually most relevant parts of the spectrum, a lower prediction order can be used in coding and synthesis to achieve the same perceptual quality as in standard linear prediction [20] [12].

In speech recognition applications, low order feature vectors are desirable because of the effect known in general pattern recognition as the curse of dimensionality [3]; basically, it means that the demand for a large number of training samples grows very rapidly with the dimensionality of the feature space. Another reason for using auditorily motivated spectrum modeling in speech recognition front-ends is that the warped frequency scale has been found to give less speaker-dependent feature representations, as was found for the PLP technique [9]. Although WLP only implements the warping of the frequency axis and none of the other perceptual aspects found in e.g. PLP, the auditory warping alone is a desirable effect.

Because the optimal prediction order depends on the sampling rate, low-order spectrum models concentrating on the perceptually more relevant areas of the spectrum could, to some extent, be achieved also by simply downsampling the signal prior to the analysis. This way, however, all of the information from the higher frequency bands is lost. Warping partially preserves this information.

6.1 Compressing formant information

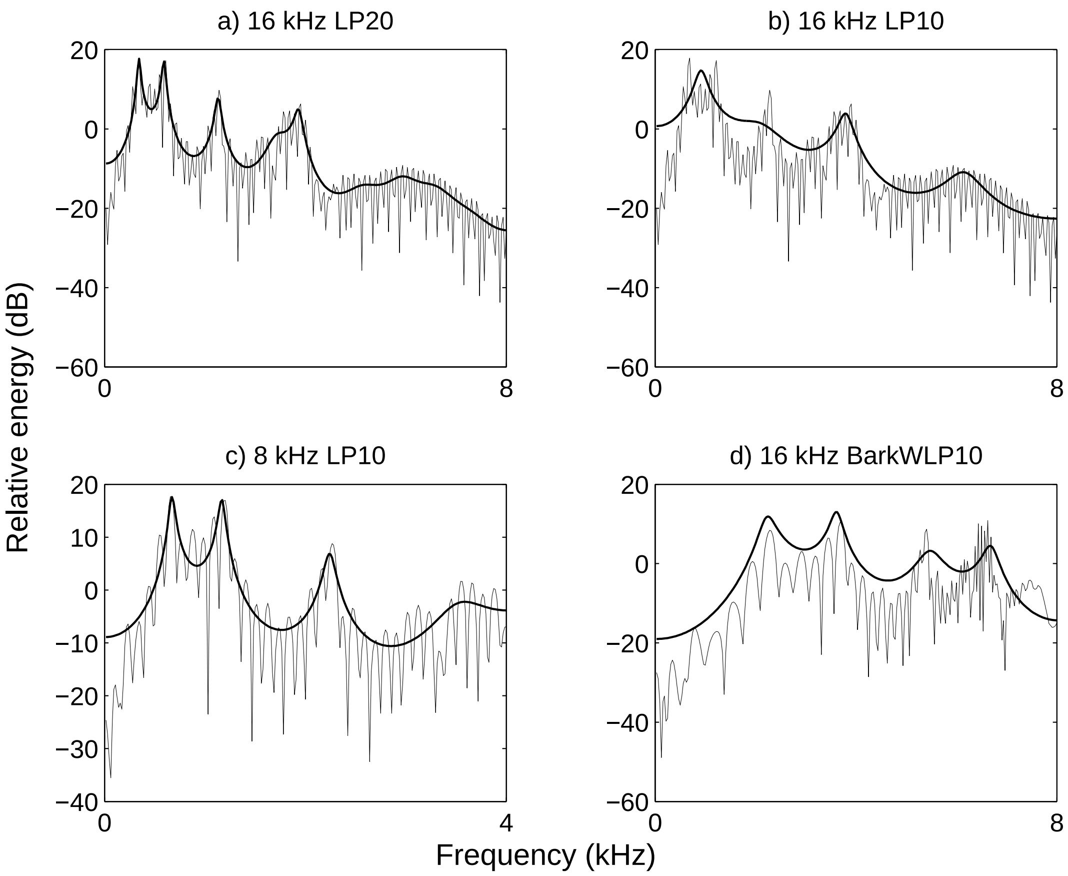

Figure 6.1: Effect of reducing the number of parameters in spectrum models generated by LP and Bark-WLP for vowel /a/ sampled at 16 kHz.

Figures 6.1 a) and 6.2 a) show ordinary linear predictive spectral estimates for vowels /a/ and /i/, respectively. The 16-kHz sampled signal is modeled using ordinary autocorrelation LP with prediction order 20. The spectral models, consisting of 20 filter coefficients, follow nicely the spectral envelopes and show the several formants within the signal bandwidth.

Suppose we want to decrease the number of model coefficients from 20 to, say, 10 while

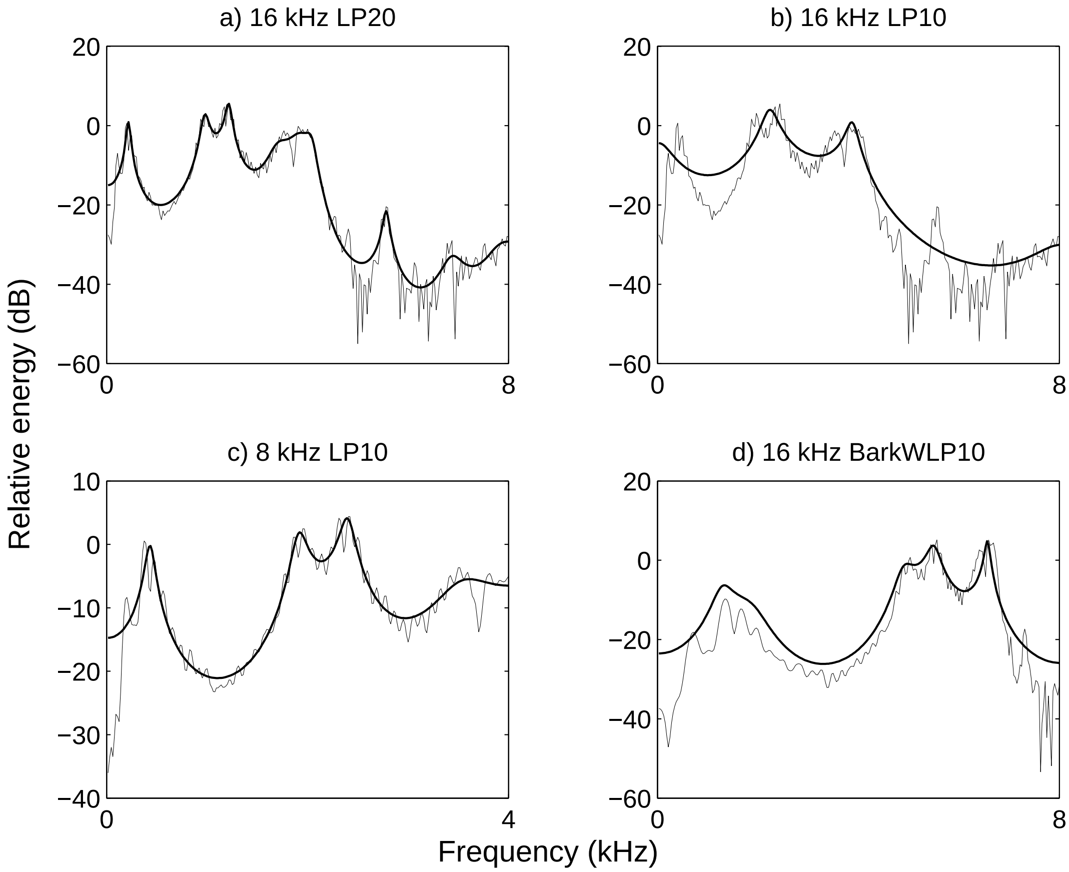

Figure 6.2: Effect of reducing the number of parameters in spectrum models generated by LP and Bark-WLP for vowel /i/ sampled at 16 kHz.

still retaining the formant structure information. Figures 6.1 b) and 6.2 b) show the spectral estimates when the number of poles in the LP model is decreased from 20 to 10. The model quality becomes very poor; some important formants are ignored altogether, while others are lumped together. This is obviously not a very good way to compress formant information by using fewer parameters.

Next, we consider downsampling the signal to 8 kHz; since by the order selection rule given in the introduction to this chapter, prediction order 10 can be considered suitable for this sampling rate. Figures 6.1 c) and 6.2 c) show the resulting spectral estimates. The lowest three formants are very well represented, but all information in the spectrum above 4 kHz is, of course, totally lost.

We then consider modeling the original 16-kHz sampled signal using autocorrelation method of Bark-warped linear prediction with order 10. The results are shown in figures 6.1 d) and 6.2 d). The warped spectrum makes no serious mistakes as case b), yet seems to capture the information of formant locations over a wide frequency band better than case c).

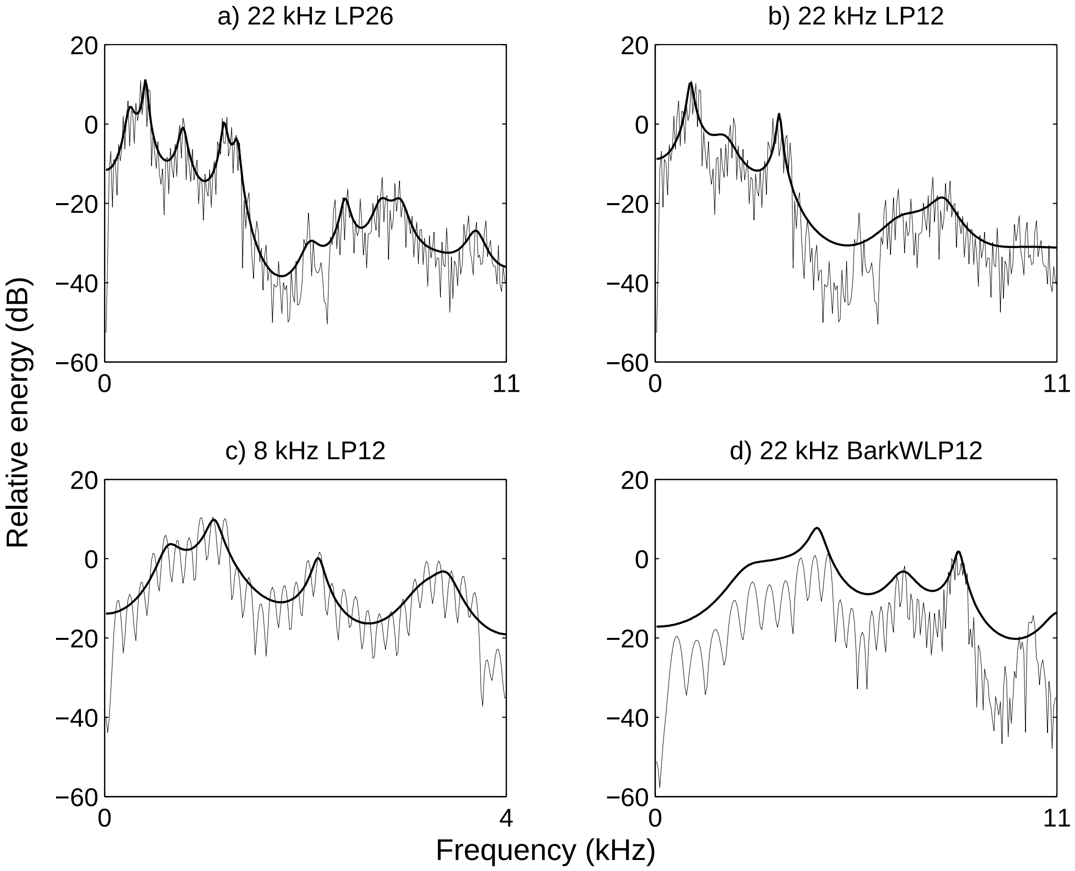

Figure 6.3: Effect of reducing the number of parameters in spectrum models generated by LP and Bark-WLP for vowel /a/ sampled at 22 kHz.

Figures 6.3 and 6.4 show similar results when a sampling rate of 22 kHz is used.

Figure 6.5 shows the vowels /a/ and /i/, sampled at 16 kHz, modeled using ordinary LP and Bark-WLP both with order 20. Using such a high prediction order the WLP models are so accurate in low frequencies that they start to depict harmonics of the fundamental frequency. Such models are probably too accurate for some applications, such as feature generation for speech recognition.

The presented examples illustrate some characteristics of Bark-WLP as well as the importance of proper order selection. The latter is discussed in the following section.

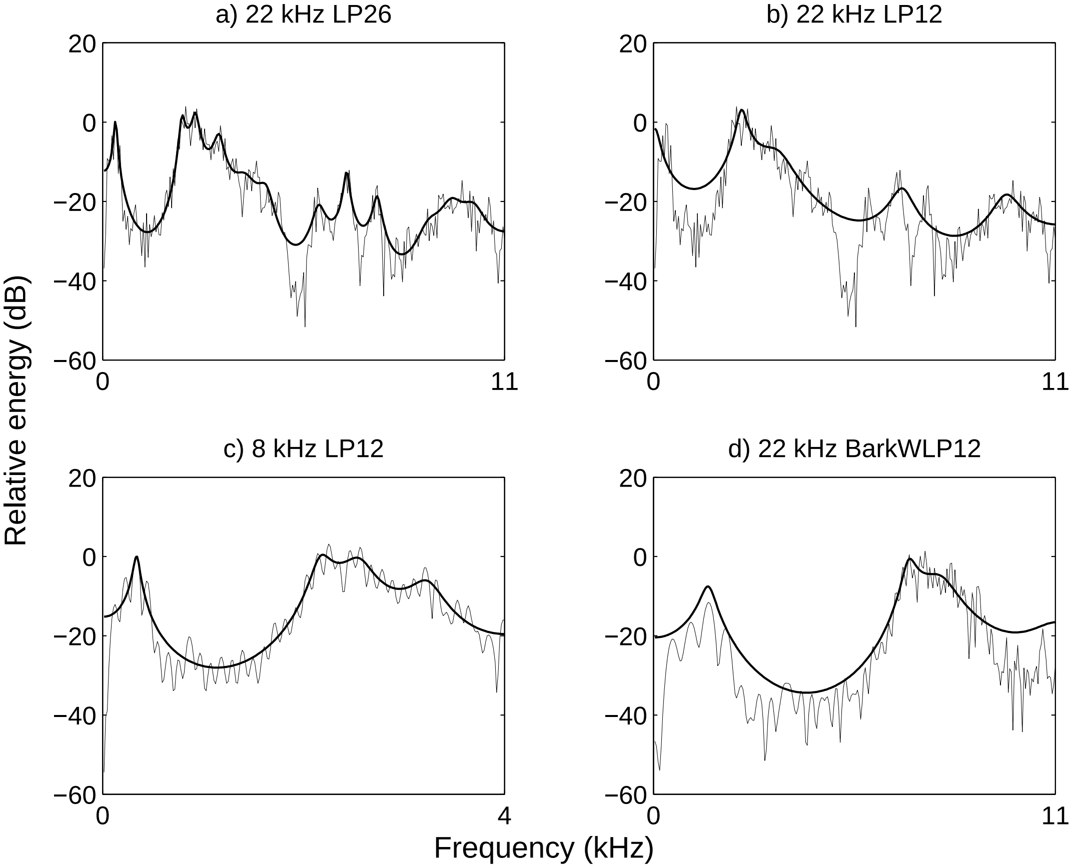

Figure 6.4: Effect of reducing the number of parameters in spectrum models generated by LP and Bark-WLP for vowel /i/ sampled at 22 kHz.

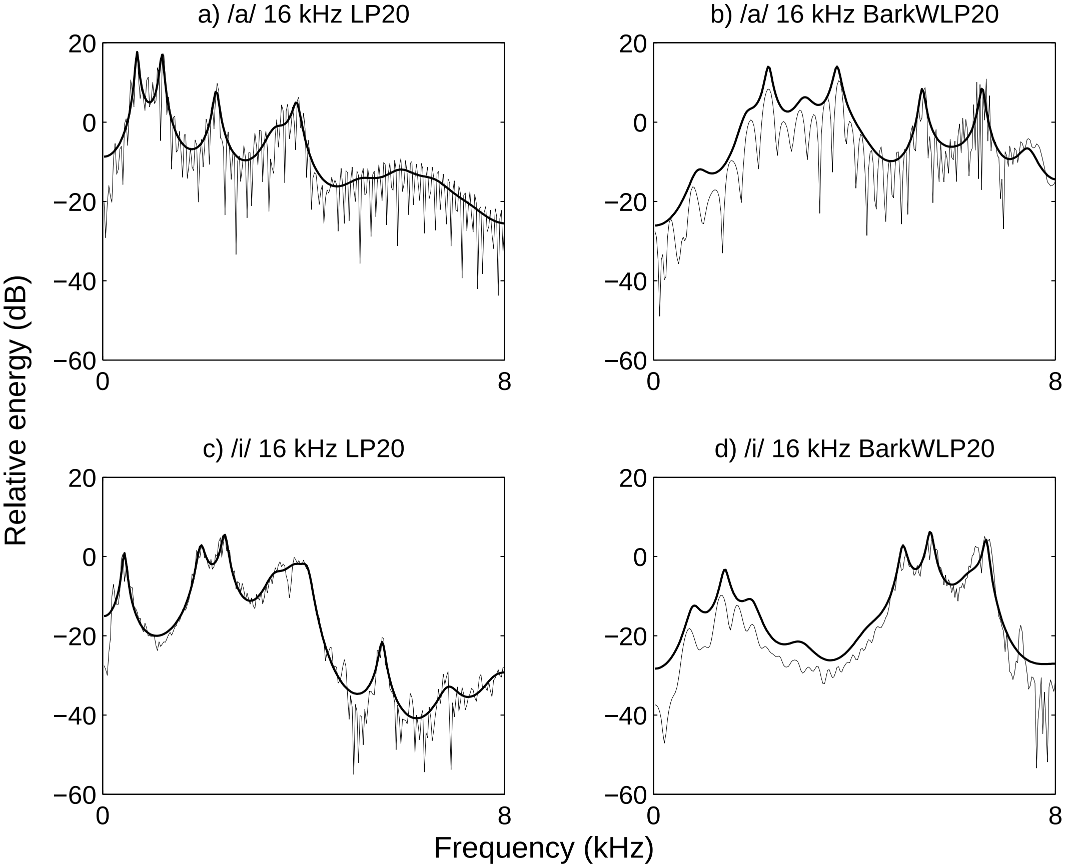

Figure 6.5: Spectrum models of order 20 generated by LP and Bark-WLP for vowels /a/ and /i/ sampled at 16 kHz.

6.2 Selection of the prediction order

Modeling performance of Bark-WLP with varying prediction order was analyzed. The goal was to determine some guidelines for the selection of prediction order in Bark-WLP. The speech material of 10 utterances, sampled at 22 kHz, was processed in frames of 25 milliseconds using a frame shift interval of 3 milliseconds. The autocorrelation formulation was used in both LP and Bark-WLP for obtaining the ordinary and warped models. For each frame, the inverse filter was applied to the original Hamming-windowed signal frame in order to obtain the residual, from which the five performance measures were computed. For each value of the prediction order, the results were averaged over some very broad phonetic class based on the manual segmentation and labeling.

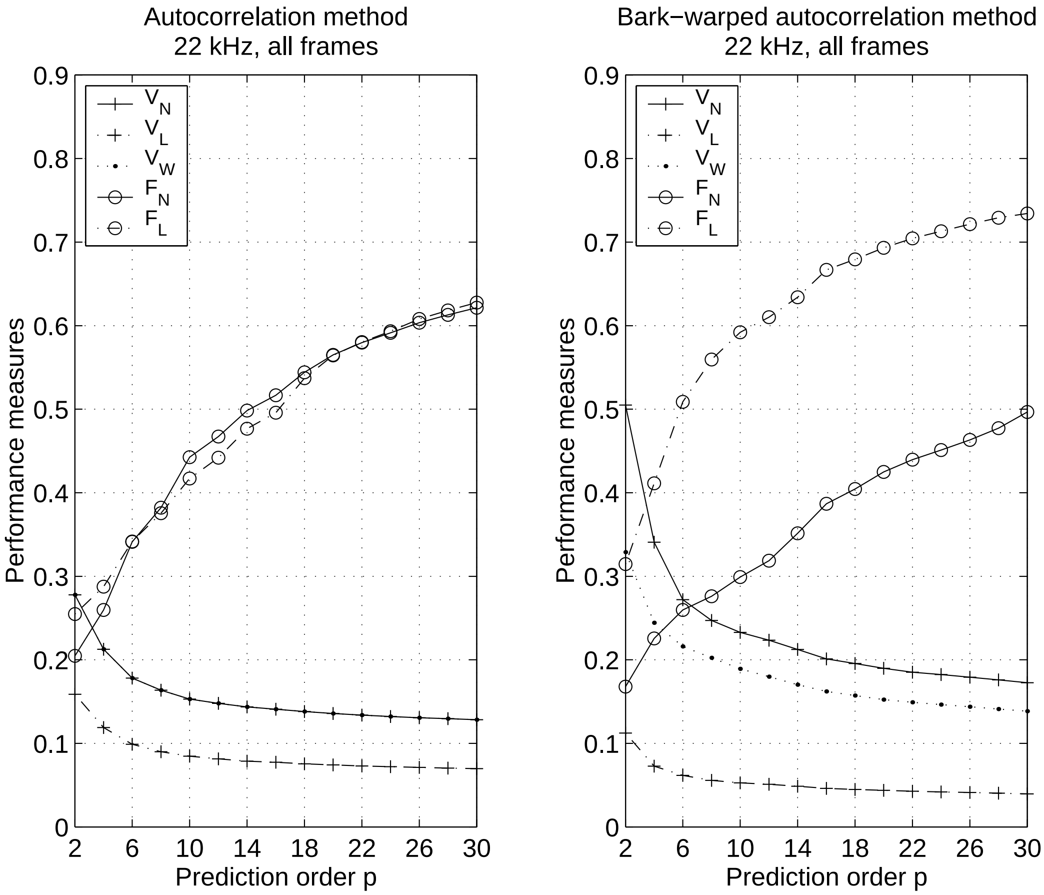

Figure 6.6: Comparison of ordinary LP and Bark-WLP, using the autocorrelation method, in terms of five performance measures averaged over all frames in speech sampled at 22 kHz.

The left pane in figure 6.6 shows the averaged performance measure curves for normal

LP when the prediction order was varied from 2 to 30. In this case the values were averaged over all frames in the material. The right pane shows similar curves for Bark-WLP; the only measures whose curves improve when moving from LP to Bark-WLP are V_L and F_L. The full-band measures V_N, V_W, F_N, are all made worse when the processing is warped. Also recall that the low-band-limited measures V_L and F_L can be expected to be maximally favorable to WLP because the frequencies considered are exactly those frequencies for which WLP improves the modeling (up to the turning point frequency).

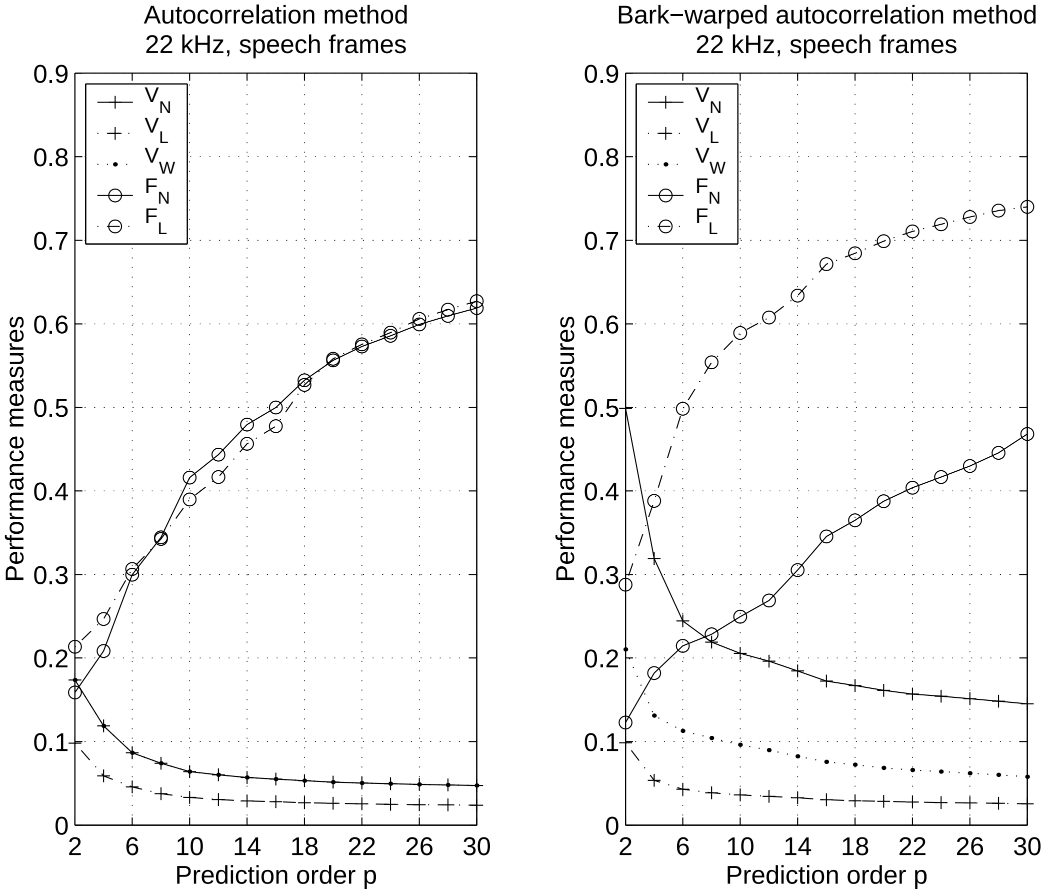

Figure 6.7: Comparison of ordinary LP and Bark-WLP, using the autocorrelation method, in terms of five performance measures averaged over non-silent frames (containing speech sounds) in speech sampled at 22 kHz.

Figure 6.7 shows similar curves when the measures are averaged only over frames containing speech; that is, the silent frames are excluded. In this case, four of the five curves show deteriorating performance when Bark-WLP is used instead of LP. However, the remaining F_L measure shows significant improvement in performance. Figure 6.8 shows

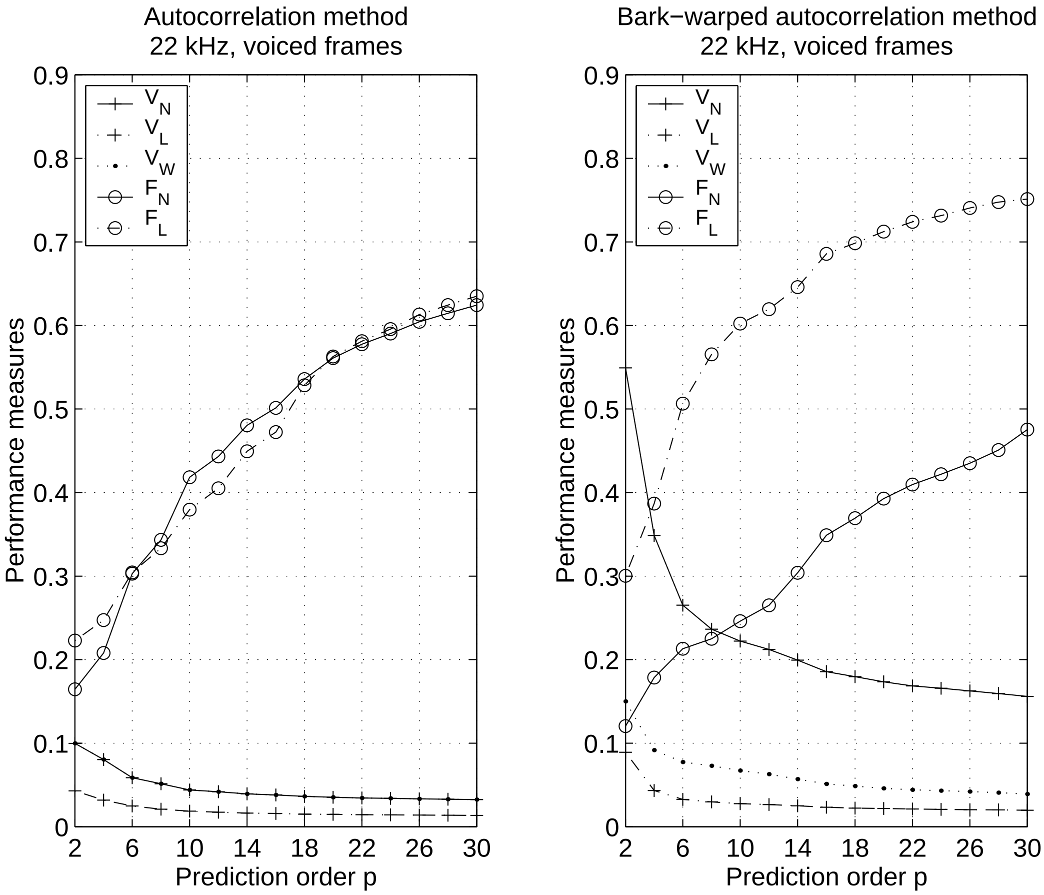

Figure 6.8: Comparison of ordinary LP and Bark-WLP, using the autocorrelation method, in terms of five performance measures averaged over voiced frames in speech sampled at 22 kHz.

very similar results when the averaging is done only over frames containing voiced speech. By these results it seems apparent that WLP improves the frequency-domain modeling in low frequencies as expected. The higher frequencies are accordingly modeled much more poorly. It seems more difficult to notice the benefits of Bark warping in the time domain, using residual energy measures, than in the spectral domain using residual whiteness measures.

It is possible to determine a value for the Bark-WLP prediction order for which the models are as accurate in low frequencies as those generated by ordinary LP using the appropriate prediction order (recall the order selection rule in the beginning of this chapter). This can be done by examining the F_L measure in both panes of each of the figures 6.6-6.8 as follows. First, select the appropriate prediction order from the left pane by the order selection rule; in this case, 24 to 26 is a good choice. Second, find the corresponding value of F_L

which in this case is approximately 0.6 in all three cases. Third, follow a straight horizontal line to the right pane until the horizontal line crosses the F_L curve. Finally, determine the Bark-WLP prediction order as the prediction order in the right pane for which the F_L measure assumes approximately the same values as F_L in the left pane using the predetermined prediction order. It can be concluded that, for a 22 kHz sampling rate, Bark-WLP with prediction order 10 to 12 gives similar spectral accuracy at signal frequencies $0 - 3048$ Hz as does ordinary LP using prediction order 24 to 26. Thus, given that we are by far most interested in these low frequencies, a significant reduction in the number of parameters in the spectral models can be achieved. However, further order reduction is not easily justified.

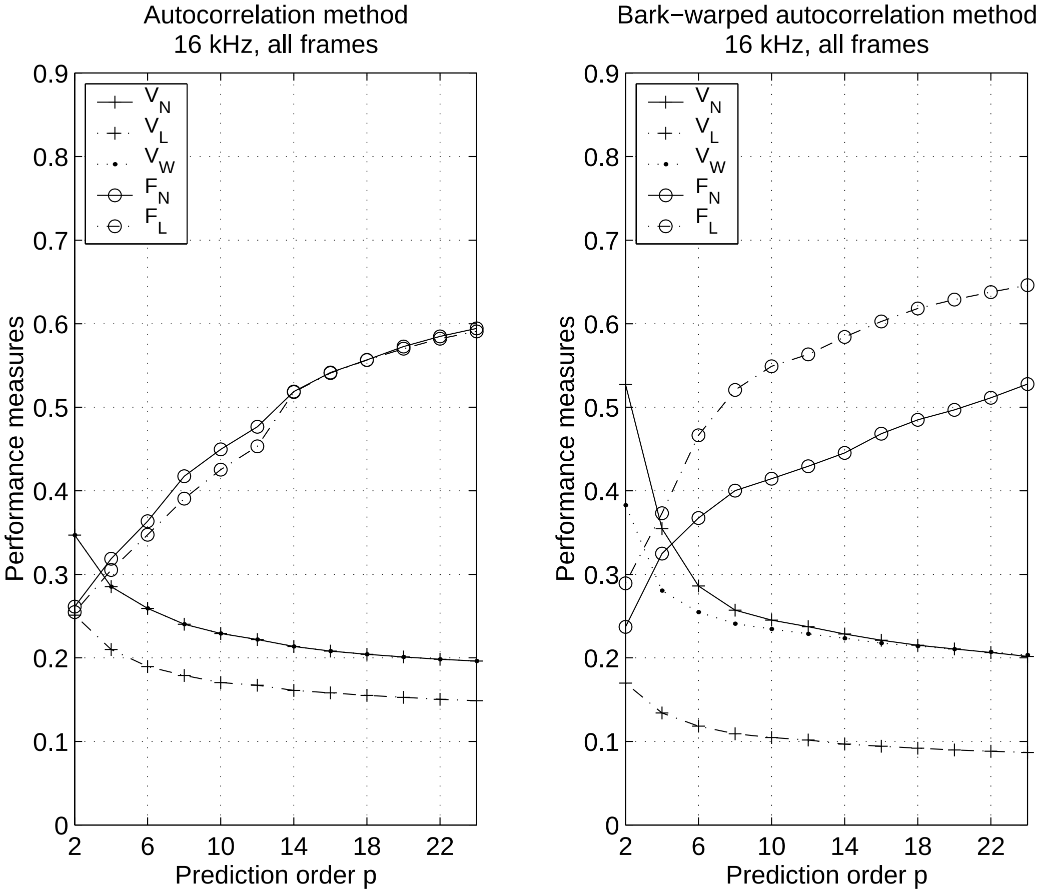

Figure 6.9: Comparison of ordinary LP and Bark-WLP, using the autocorrelation method, in terms of five performance measures averaged over all frames in speech sampled at 16 kHz.

Figures 6.9-6.11 show the similar curves for the same speech material downsampled to 16 kHz. In this case the turning point frequency used in the computation of V_L and F_L was 2437 Hz. Again, the low frequency band up to that frequency should be precisely the

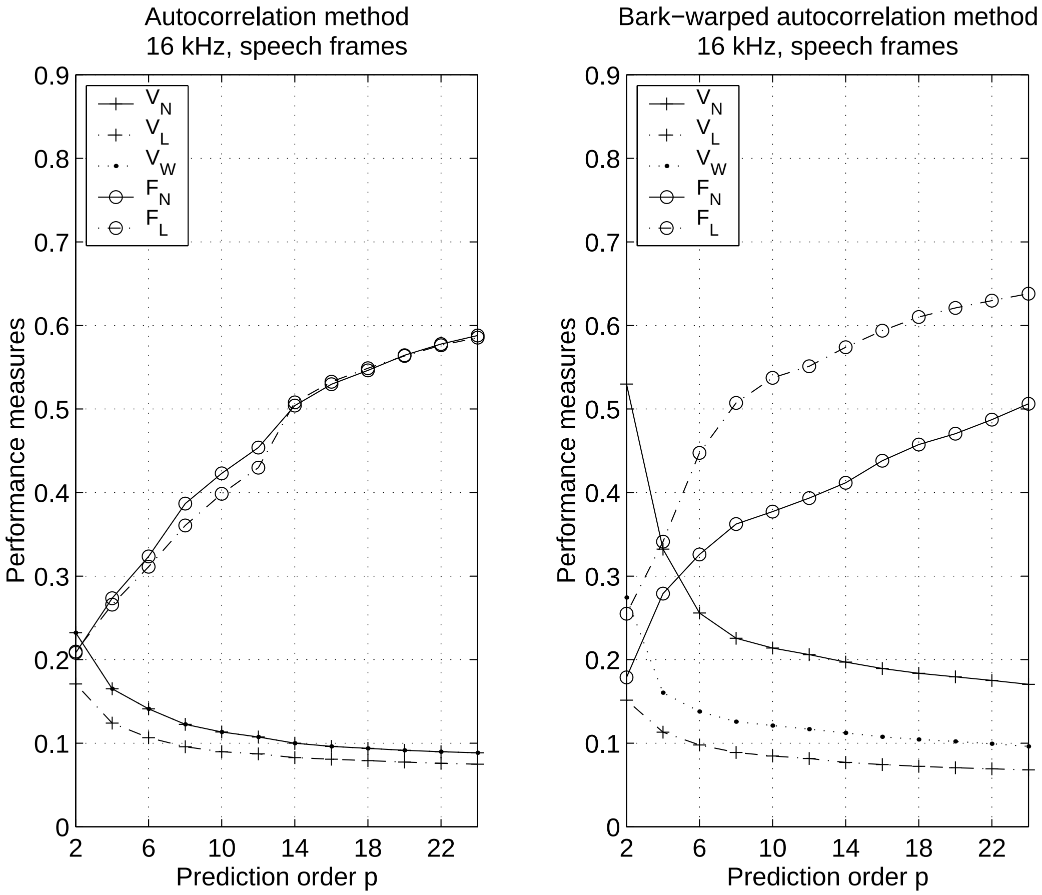

Figure 6.10: Comparison of ordinary LP and Bark-WLP, using the autocorrelation method, in terms of five performance measures averaged over non-silent frames (containing speech sounds) in speech sampled at 16 kHz.

band for which using WLP is beneficial. Repeating the steps in determining the Bark-WLP prediction order, the ordinary LP order is first chosen as 18. This leads to choosing 12 as the corresponding Bark-WLP prediction order. Ordinary LP order 20 leads to 13 or 14 as the Bark-WLP order. It seems that here, the benefit of using Bark-WLP is less than with a 22 kHz sampling rate. This is natural considering the relation of the Bark scale to the frequency scale (figure 2.5); the more the sampling rate (and the Nyquist frequency) decreases, the larger is the proportion of the frequencies where the frequency and Bark scales are almost linearly related. In other words, if the signal is already severely bandlimited to the important low frequencies, where the critical bandwidth does not vary too much, an auditory frequency mapping can not improve the representation by emphasizing some frequencies and de-emphasizing others.

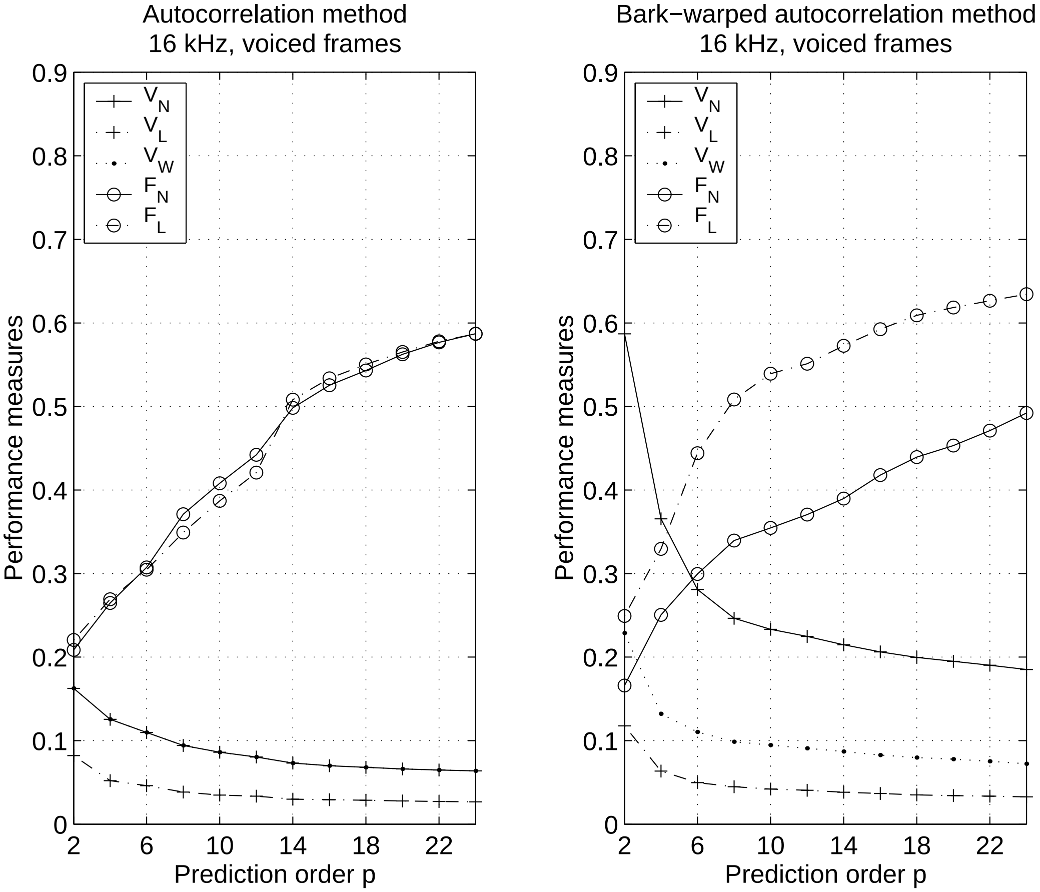

Figure 6.11: Comparison of ordinary LP and Bark-WLP, using the autocorrelation method, in terms of five performance measures averaged over voiced frames in speech sampled at 16 kHz.

Chapter 7

Adjusting the time resolution

The WLP spectrum estimation methods discussed in this thesis can be divided into two main groups: block methods (autocorrelation and covariance methods) and adaptive methods (LMS and GAL). The block methods use a fixed time window for extracting signal frames, from which the spectrum estimates are computed. For a given frame shift interval, the length of the estimation frame determines the time resolution of the resulting time-frequency representation depicted by the spectrogram. Shortening the window increases the time resolution, while lengthening the window trades off time resolution for more stable and robust spectrum estimates.

The adaptive methods discussed in this thesis update the estimates sample by sample. The adaptation rate parameter determines the speed at which old information is replaced by new information in the estimate or, in other words, specifies how long memory the adaptive filter has. The effect of the adaptation rate on the time-frequency representation is inversely analogous to the effect of the window length in block methods; increasing the adaptation rate increases the time resolution while decreasing the adaptation rate decreases it.

When conventional block methods of linear prediction are used (pitch-asynchronously), the time resolution is usually set to 20-30 milliseconds. This avoids excessive modulation by the glottal excitation while still being fairly accurate in both time and frequency. The goal here is to determine the adaptation parameters of the adaptive methods so as to achieve a time resolution comparable with that of a block method using a known frame length. In section 7.1 a simple method is introduced for determining analysis parameter values corresponding approximately to a fixed block estimation frame length in terms of the time resolution. As will be seen in section 7.4, this method also enables us to assess the relative modeling performance of the four WLP methods. The determined parameter values are given in section 7.2. Section 7.3 includes visualizations (spectrograms) of the time-frequency representations with different time resolution characteristics.

7.1 A method for determining analysis parameters

Associated with each WLP method discussed in this thesis, there is one relevant analysis parameter that primarily controls the time resolution by determining either the block size or the adaptation rate. Each such choice may be thought to lead to some value for the length of an *effective estimation frame* (using a rectangular time window), given some modeling performance measure. The desired length of such time window may be specified by the length of an *evaluation frame*; this is just the length of the rectangular window used in extracting frames from which the five modeling performance measures are computed. Each performance measure should assume its best values when the effective estimation frame is roughly the same size as the evaluation frame.

Figure 7.1 illustrates the concept of two independent time frames. The length of the effective estimation frame is made as closely equal to the length of the evaluation frame as possible. This is done by choosing an optimal value of the analysis parameter controlling the length of the estimation frame.

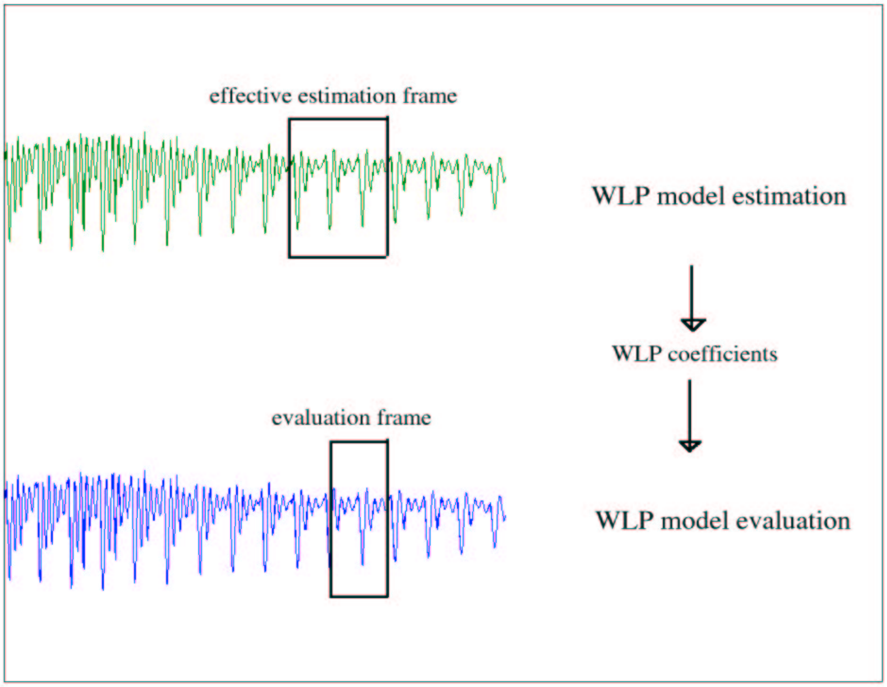

Figure 7.1: An illustration of effective model estimation frame and model evaluation frame taken from the signal.

More specifically, the WLP residual is first obtained from an evaluation frame by inverse filtering. The residual is then used, either independently or together with the original signal frame, in computing the values for the performance measures associated with the frame. The evaluation frame lengths used in these simulations are listed in table 7.1. The tested analysis parameter values are listed in table 7.2.

Table 7.1: Length of the evaluation frame in milliseconds.

15	20	25	30	35

Table 7.2: Candidate analysis parameters for four WLP methods.

AUT	COV	LMS	GAL
frame	frame	step	memory
size (ms)	size (ms)	size	coefficient
10	10	0.001	0.9795
15	15	0.005	0.9820
20	20	0.010	0.9845
25	25	0.020	0.9870
30	30	0.030	0.9895
35	35	0.040	0.9920
40	40	0.050	0.9930
45	45	0.060	0.9940
50	50	0.070	0.9950
		0.080	0.9960
		0.090	0.9970
		0.100	0.9980
		0.150	0.9990
		0.200	0.9995

For each WLP method and each evaluation frame length, the following is done: first, with the analysis parameter value fixed, the performance measures are averaged over some set of frames. Second, the minimum or maximum (whichever corresponds to the best models) of each averaged performance measure is located over all values of the analysis parameter. The found parameter values are the recommended "best" values, in order to get the time resolution given by the length of the evaluation frame, judging by the respective measures. The found minima and maxima can be used to compare the methods with each other.

7.2 Choosing the analysis parameters

In this section we determine the optimal analysis parameter for each of the four methods in order to optimize the overall average modeling accuracy (all signal frames considered). The results for the block methods are discussed first, although the optimal choice of the analysis frame size may seem self-evident. This is done in order to verify the method for choosing the optimal parameter. In addition, the autocorrelation and covariance methods differ slightly in this respect.

Figure 7.2 shows what happens with each of the five features, computed in five different time frames, when the WLP analysis frame length is varied in the Bark-WLP autocorrelation method. The best value of each curve is marked with an additional square. It is not surprising that the curves with a short evaluation frame have their best values when the WLP estimation frame is also short; conversely, a longer evaluation frame favors a long WLP estimation frame. The corresponding values for WLP estimation frame length are given in table 7.3. The values are very close to the length of the evaluation frame which was used.

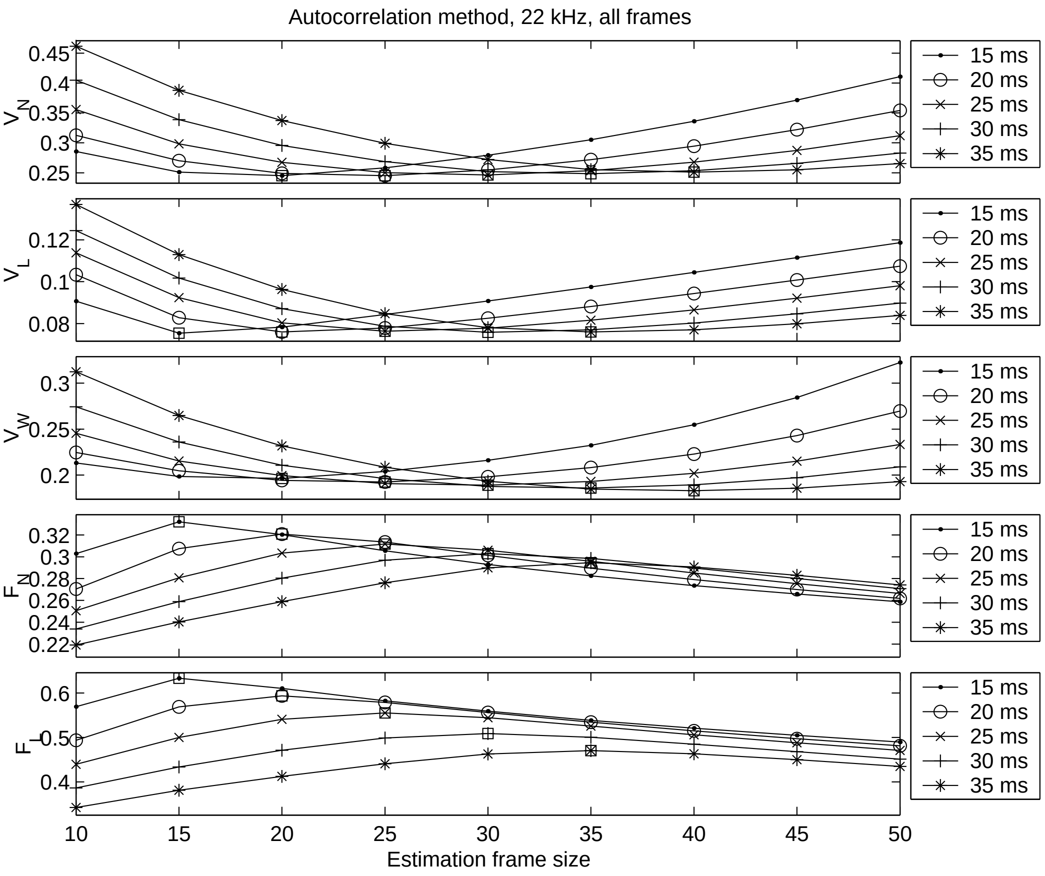

Figure 7.2: Effects of varying the estimation frame size of Bark-WLP autocorrelation method on the model evaluation measures using different-sized evaluation frames.

Table 7.3: Optimum block size (in milliseconds) for the autocorrelation method with 22 kHz sampling rate.

Measure	Evaluation frame size (ms)				
	15	20	25	30	35
V_N	20	25	30	35	40
V_L	15	20	25	30	35
V_W	20	25	30	35	40
F_N	15	20	25	30	35
F_L	15	20	25	30	35

Figure 7.3 and table 7.4 give analogous results for the covariance method of Bark-WLP. The covariance method favors a slightly shorter estimation frame, which is not surprising considering the differences in windowing.

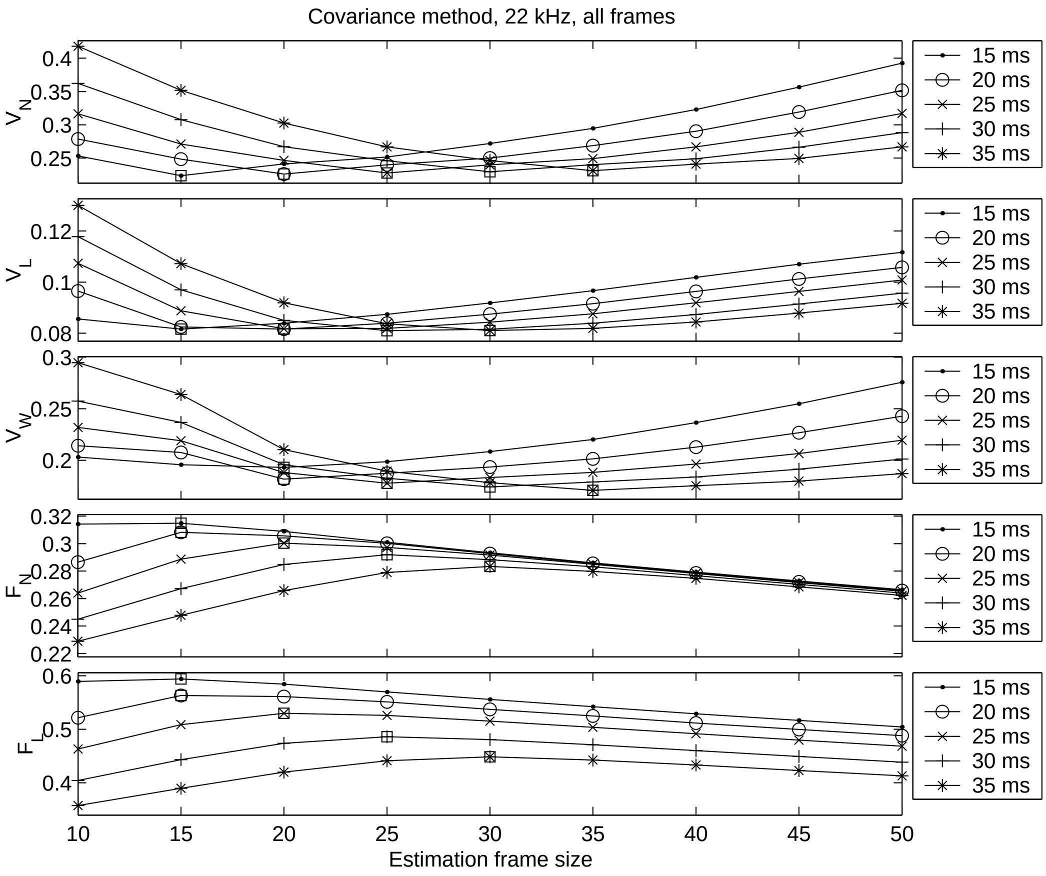

Figure 7.3: Effects of varying the estimation frame size of Bark-WLP covariance method on the model evaluation measures using different-sized evaluation frames.

Table 7.4: Optimum block size (in milliseconds) for the covariance method with 22 kHz sampling rate.

	Evaluation frame size (ms)				
Measure	15	20	25	30	35
V_N	15	20	25	30	35
V_L	15	20	20	25	30
V_W	20	20	25	30	35
F_N	15	15	20	25	30
F_L	15	15	20	25	30

Figure 7.4 and table 7.5 show the results for the GAL method. For e.g. 25 ms resolution, the memory parameter could be set to 0.993. The effective time window length is quite sensitive to the value of the parameter. It should be observed that while the optimum values

for the memory parameter are quite close to unity, it should not be set too near to unity (meaning too slow adaptation) or the performance may degrade radically.

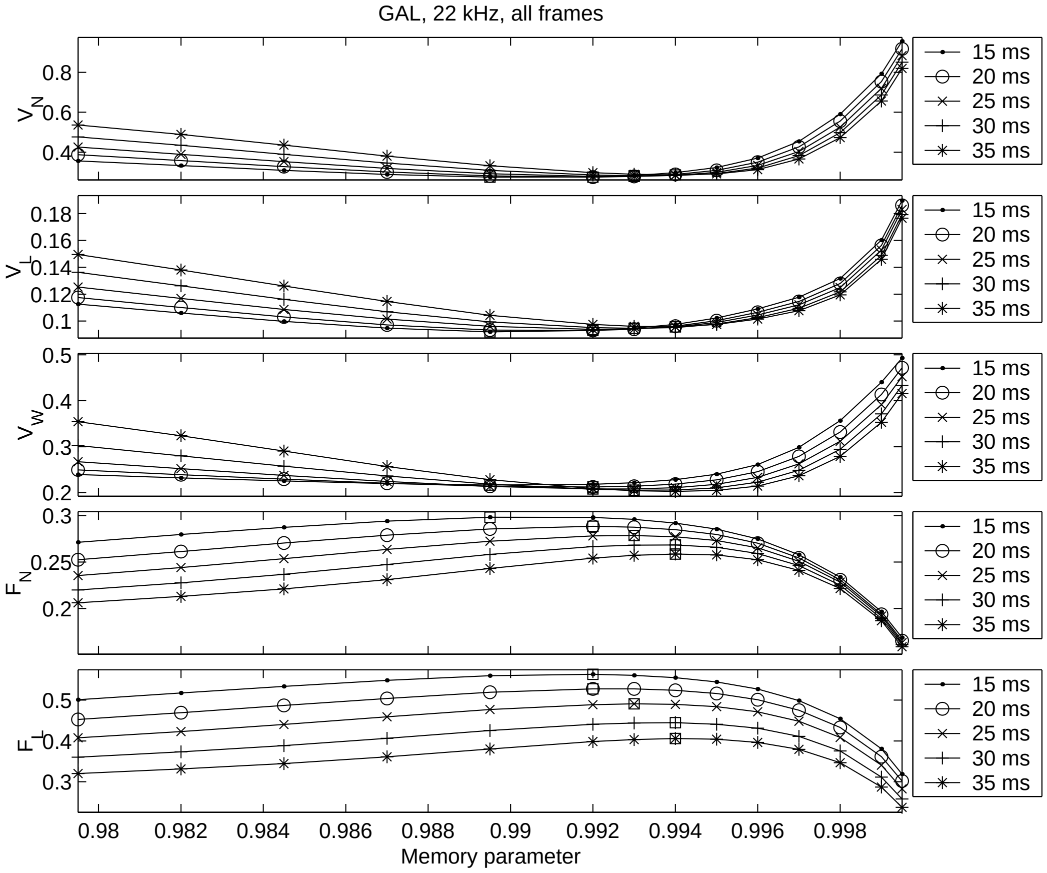

Figure 7.4: Effects of varying the memory parameter of the Bark-WLP GAL method on the model evaluation measures using different-sized evaluation frames.

Table 7.5: Optimum memory parameter for the GAL method with 22 kHz sampling rate.

	Evaluation frame size (ms)				
Measure	15	20	25	30	35
V_N	0.9895	0.9920	0.9930	0.9930	0.9940
V_L	0.9895	0.9920	0.9920	0.9930	0.9940
V_W	0.9895	0.9920	0.9920	0.9930	0.9940
F_N	0.9895	0.9920	0.9930	0.9940	0.9940
F_L	0.9920	0.9920	0.9930	0.9940	0.9940

Finally, figure 7.5 and table 7.6 show the results for LMS. The curves resemble the inverted versions of the GAL curves. While the optimum value for the update step size parameter is quite small, it must at the same time be large enough to allow the filter to track the time-varying signal statistics. For a 25 ms window, the step size could be set to 0.06.

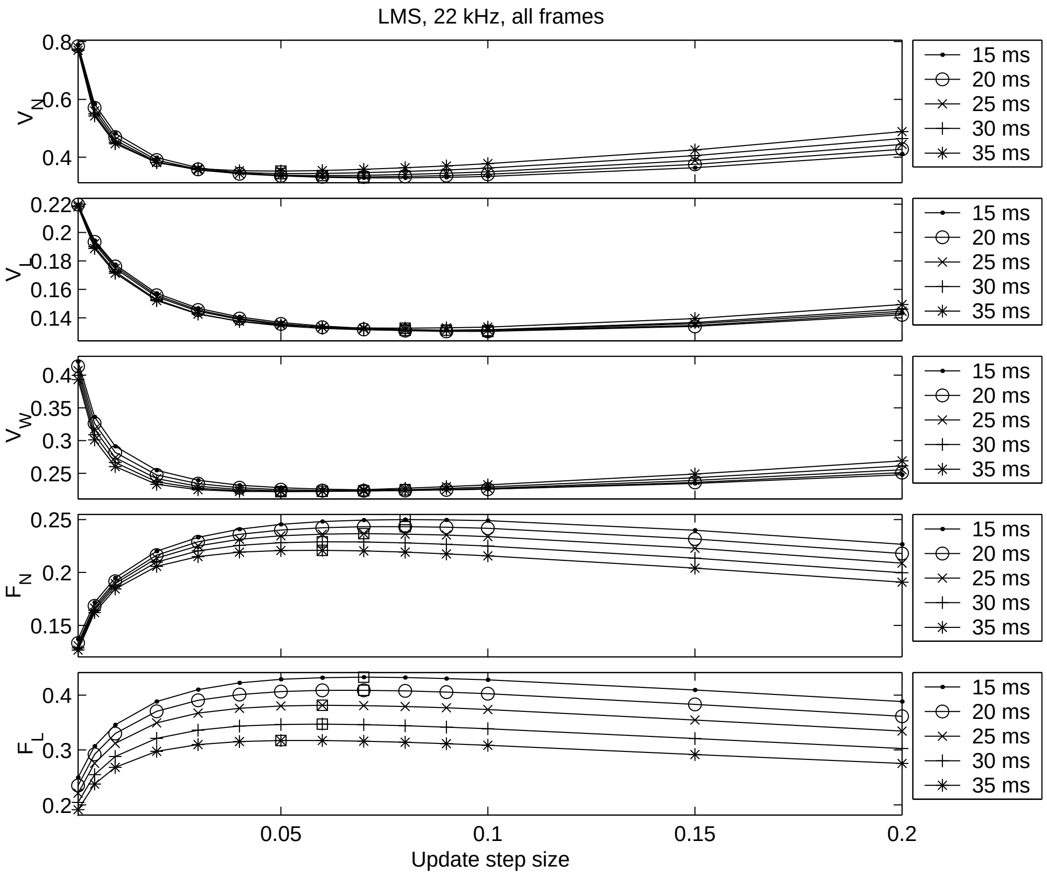

Figure 7.5: Effects of varying the update step size parameter of Bark-WLP LMS method on the model evaluation measures using different-sized evaluation frames.

Table 7.6: Optimum update step size for the LMS method with 22 kHz sampling rate.

	Evaluation frame size (ms)				
Measure	15	20	25	30	35
V_N	0.07	0.07	0.06	0.05	0.05
V_L	0.10	0.10	0.09	0.08	0.08
V_W	0.08	0.07	0.06	0.05	0.05
F_N	0.08	0.08	0.07	0.06	0.06
F_L	0.07	0.07	0.06	0.06	0.05

For reference, tables 7.7 and 7.8 show the optimized adaptation rate parameters for GAL and LMS, respectively, when the sampling rate is 16 kHz.

Table 7.7: Optimum memory parameter for the GAL method with 16 kHz sampling rate.

	Evaluation frame size (ms)				
Measure	15	20	25	30	35
V_N	0.9895	0.9895	0.9920	0.9930	0.9930
V_L	0.9895	0.9895	0.9920	0.9920	0.9930
V_W	0.9895	0.9895	0.9920	0.9930	0.9940
F_N	0.9895	0.9920	0.9920	0.9930	0.9940
F_L	0.9895	0.9920	0.9920	0.9930	0.9930

Table 7.8: Optimum update step size for the LMS method with 16 kHz sampling rate.

	Evaluation frame size (ms)				
Measure	15	20	25	30	35
V_N	0.08	0.07	0.06	0.05	0.04
V_L	0.07	0.07	0.06	0.06	0.05
V_W	0.08	0.07	0.06	0.05	0.04
F_N	0.10	0.08	0.07	0.07	0.06
F_L	0.08	0.08	0.07	0.07	0.06

7.3 Example spectrograms

This section shows spectrograms generated by the WLP techniques. First, figure 7.6 shows what may happen if the analysis methods are not tuned properly. Each spectrogram depicts the same utterance of the Finnish sentence ’Väärä ksylofonivirtuoosi nuutuu häätölaeissa’. Panes a) and b) show the results of the autocorrelation method with 50 ms and 10 ms estimation frame, respectively. The former is fairly good in voiced speech, but may smear or distort the phone boundaries with abrupt changes. On the other hand, the 10 ms analysis results in some noise in the spectrogram. Panes c) and d) of figure 7.6 show the GAL method with two adaptation parameters 0.9995 and 0.9795. The former leads to extremely slow adaptation and a spectrogram that is totally unacceptable for most practical purposes. The latter leads to fast adaptation and produces a spectrogram that is even more noisy than the one in pane b).

Spectrograms for the same utterance produced by methods optimized for 15 ms, 25 ms, and 35 ms frames are shown in figures 7.7, 7.8, and 7.9, respectively. Each figure shows the spectrograms obtained after tuning each of the four methods to best approximate the frame size in question. Overall, there are not very big differences between the block methods and GAL. LMS is of poor quality also by this visual inspection. There are also not very big differences between 15 ms and 35 ms resolutions; however, the former is somewhat more noisy as could be expected. The 15 ms frame may be a bit too short for high quality speech analysis. Also classic LP literature says that, at least for the autocorrelation method, the estimation frame should contain at least two pitch periods.

A close look at the spectrograms shows that GAL may have minor problems in rapid transitions from speech to silence; the silence regions tend to be narrower than in the autocorrelation and covariance methods, even though the adaptation rate is optimized. Especially the low formants may be sustained in the models longer than they should. This behavior is apparent especially with slower adaptation rates (figures 7.8 and 7.9). A natural explanation of the phenomenon is that speech contains both abrupt and gradual changes in both the excitation source and the vocal tract filter, and it may be impossible to capture all the transitions perfectly with a constant adaptation rate. The block methods, on the other hand, have no memory beyond the analysis frame and can immediately “adapt” to new conditions as soon as the analysis frame size allows. A potential problem associated with this observed behavior of the GAL method is that it distorts the phone boundaries and may make the occlusion parts of stop consonants too short for reliable recognition.

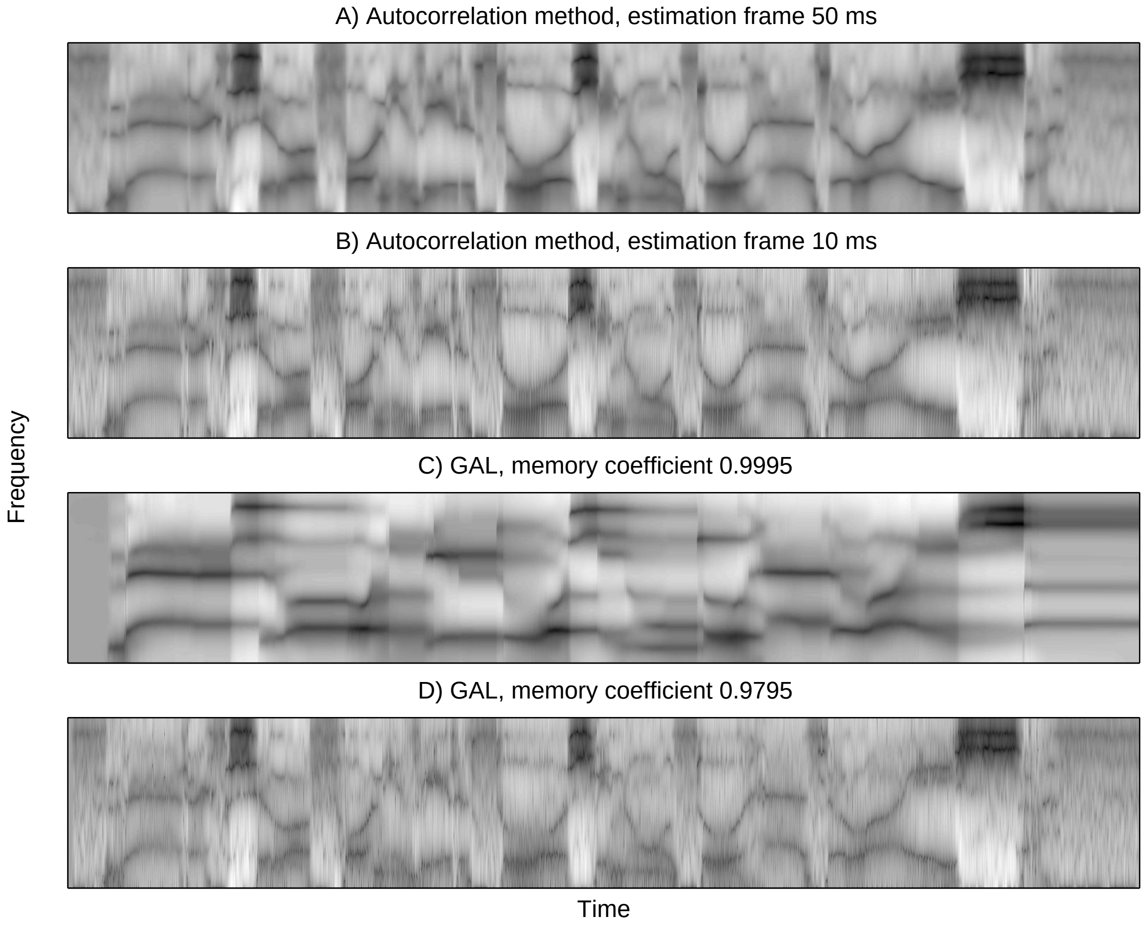

Figure 7.6: WLP spectrograms generated by analysis methods with non-optimal tuning for phone separation: A) autocorrelation method with a long 50 ms estimation frame, B) autocorrelation method with a short 10 ms estimation frame, C) GAL method with a very slow adaptation rate, D) GAL method with fast adaptation.

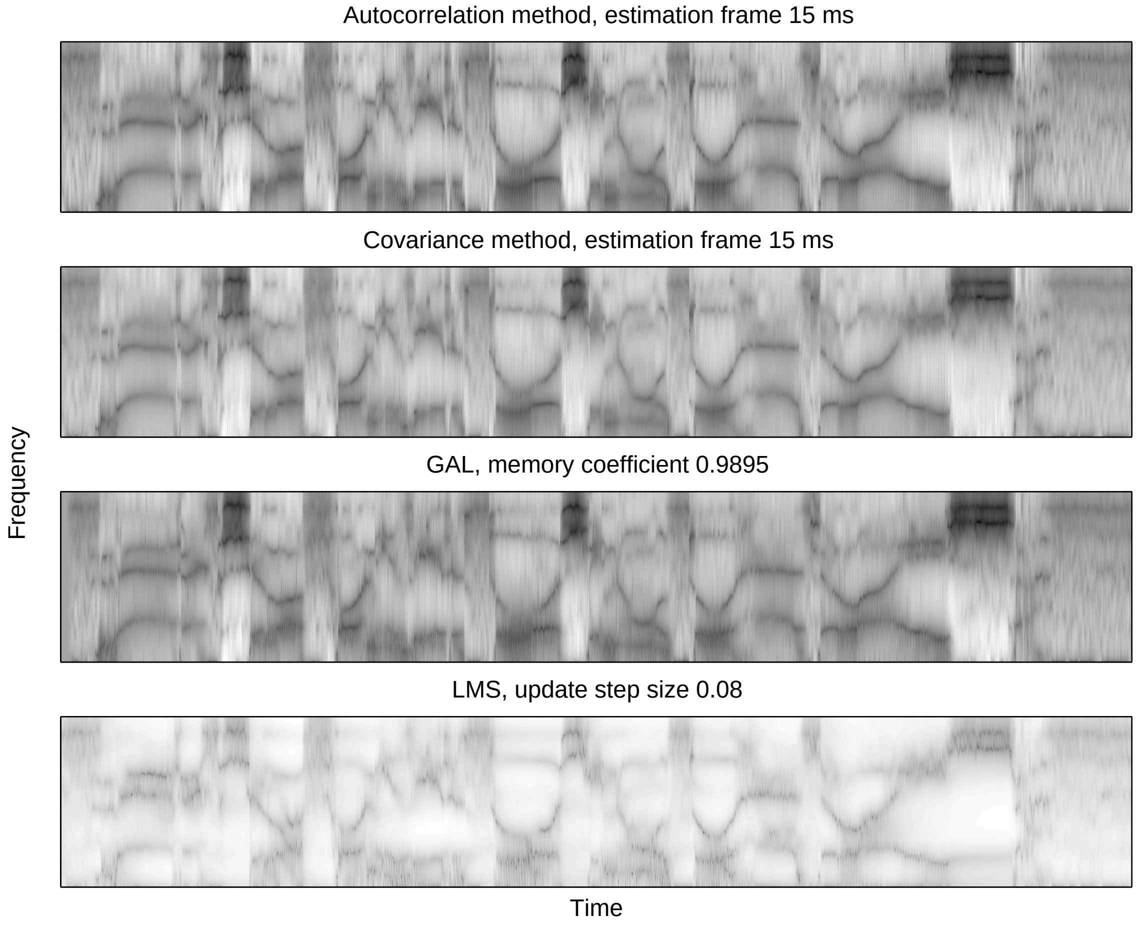

Figure 7.7: WLP spectrograms generated by autocorrelation, covariance, GAL, and LMS after optimal tuning of the methods to approximate a 15 ms time frame.

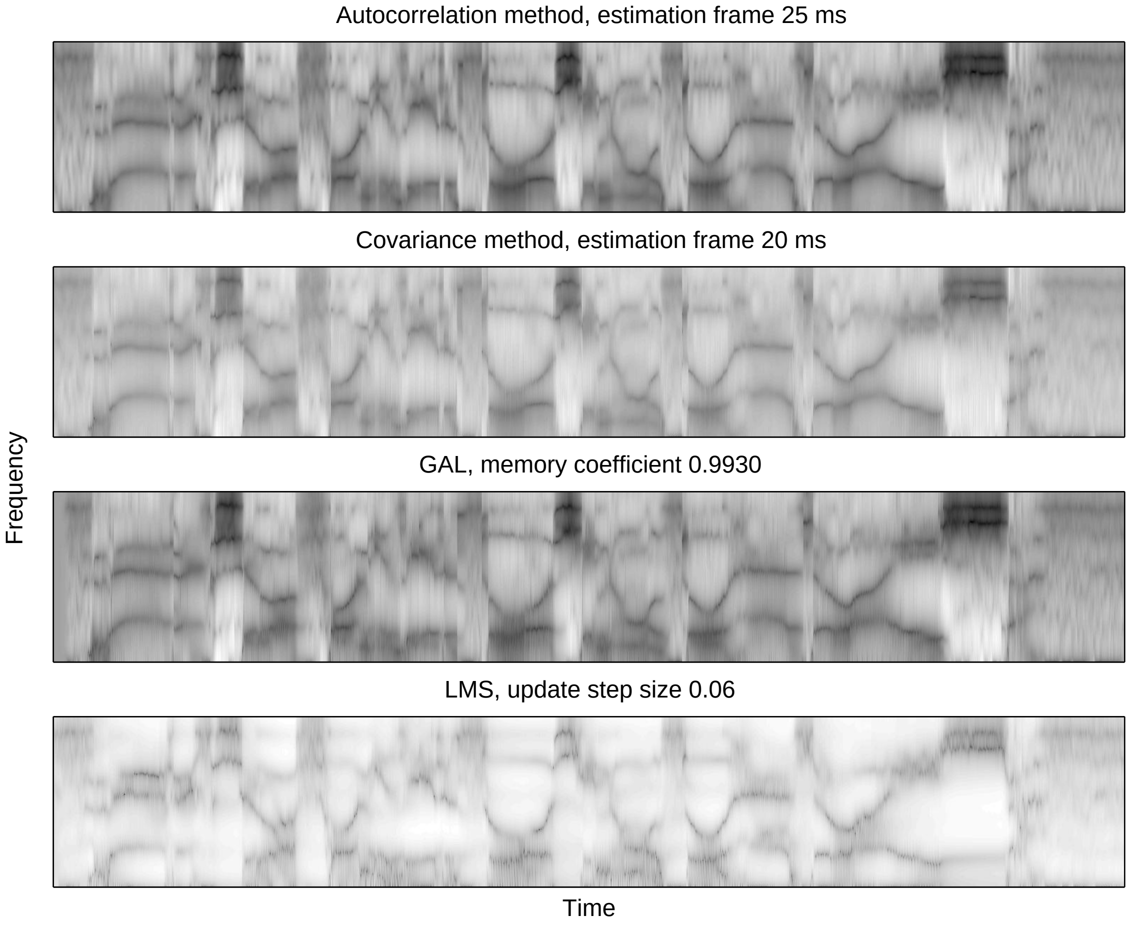

Figure 7.8: WLP spectrograms generated by autocorrelation, covariance, GAL, and LMS after optimal tuning of the methods to approximate a 25 ms time frame.

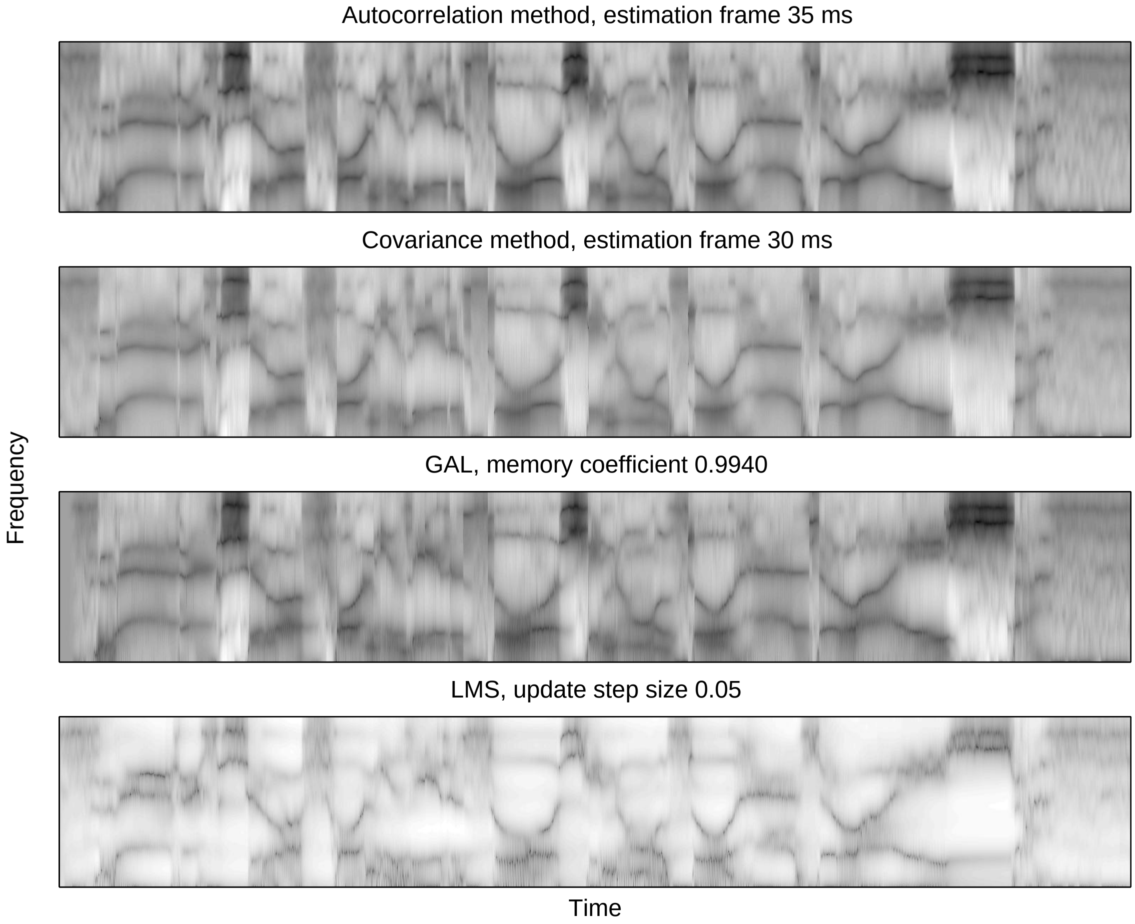

Figure 7.9: WLP spectrograms generated by autocorrelation, covariance, GAL, and LMS after optimal tuning of the methods to approximate a 35 ms time frame.

7.4 Comparison of WLP techniques

In this last section the viewpoint is that of comparing the relative modeling performance of the four methods after they have been optimized for model evaluation windows of different lengths. The results were averaged over different broad phonetic categories. The results are represented graphically in figures 7.10-7.14, with each figure representing one of the five performance measures. As was seen in chapter 6, each performance measure depicts different features of the models' fit to the data.

Each pane in each figure represents averaging over different sets of speech sounds. *All frames* and *speech frames* are the most difficult categories for the adaptive methods, since these sound categories may include rapid transients during e.g. stop consonants. *Voiced frames* includes sudden changes in the vocal tract filter (e.g. sudden opening of the main vocal tract at the end of a nasal consonant). *Vowel frames* is the easiest to track since it contains no abrupt changes, just slow gradual changes in the formant frequencies.

All five performance measures tell more or less the same. When analyzed this way, the block methods are better than the adaptive methods. LMS is the worst in all cases, as could be expected. GAL is slightly less accurate when all frames are considered, but comes close to the block methods when we focus on the more stationary parts of the speech signal. Apparently, the slow adaptation required for optimally small misadjustment during these stationary segments is perhaps too slow for rapid changes occurring e.g. with stop consonants. Not surprisingly, the results seem to support a claim that a simple adaptive gradient method, such as GAL, can not achieve a total modeling performance exactly comparable to the block methods; at least not using a constant adaptation rate.

The normalized residual energy measures V_N, V_W, and V_L achieve similar minimum values in evaluation windows of different lengths. The residual whiteness measures F_N and F_L, when maximized with respect to the analysis parameter, show better modeling performance in shorter time windows. This behavior is especially apparent with F_L, when only the lowest frequencies are considered.

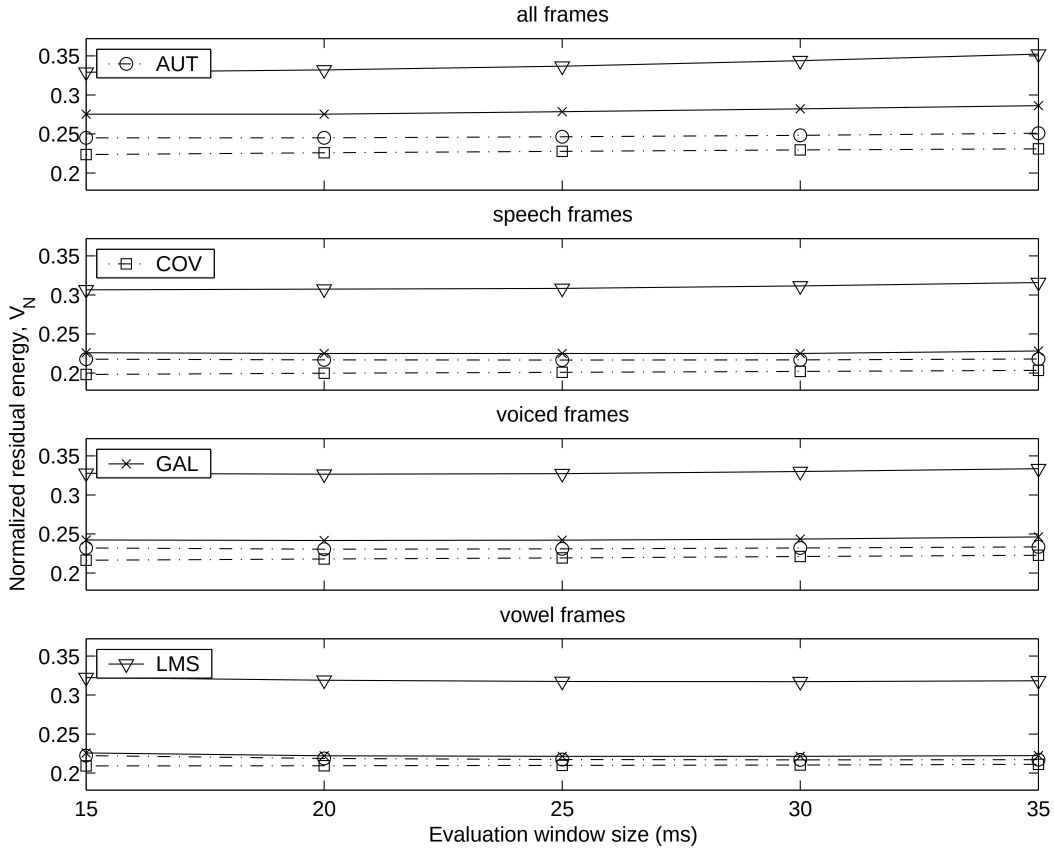

Figure 7.10: Best achieved average V_N, as a function of the evaluation frame length, for four phonetic categories.

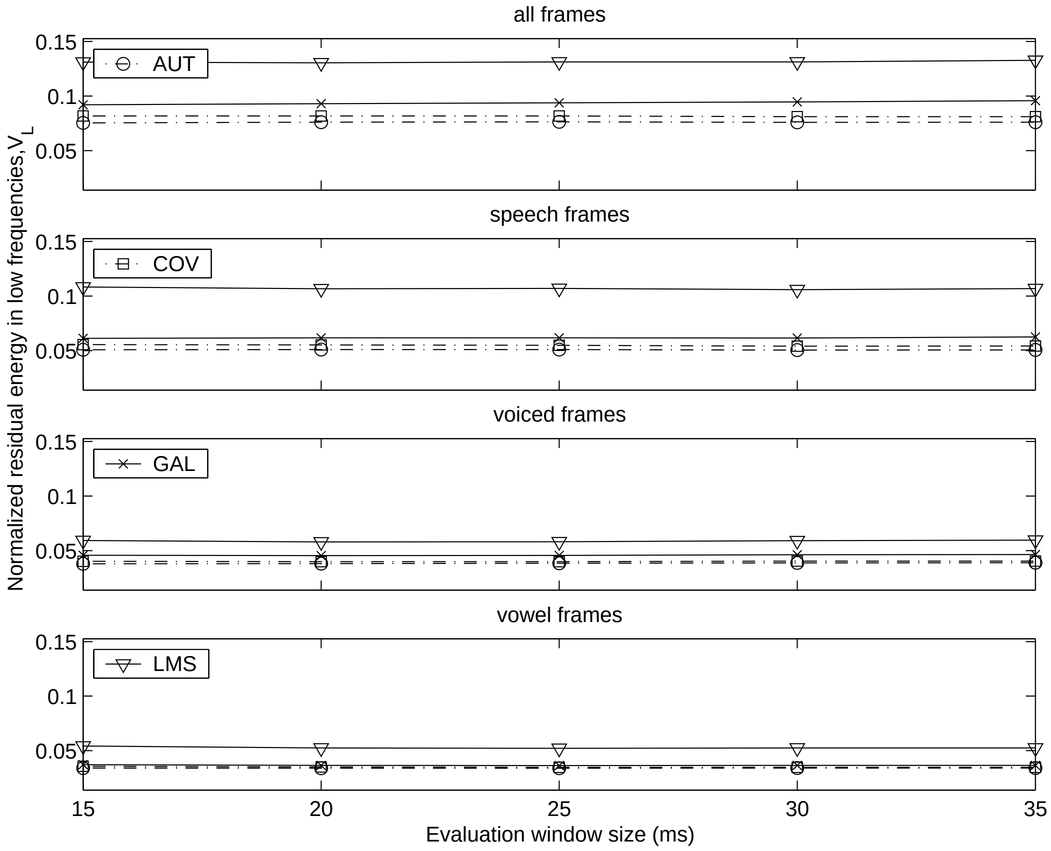

Figure 7.11: Best achieved average V_L, as a function of the evaluation frame length, for four phonetic categories.

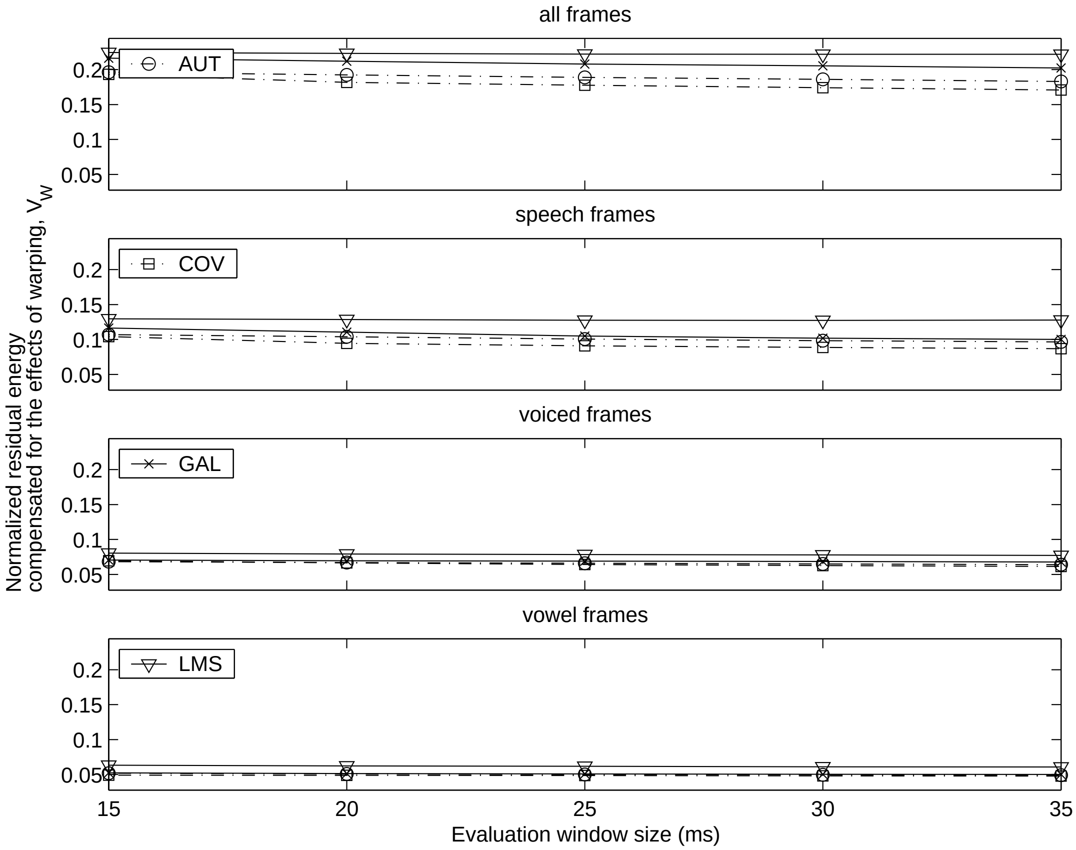

Figure 7.12: Best achieved average V_W, as a function of the evaluation frame length, for four phonetic categories.

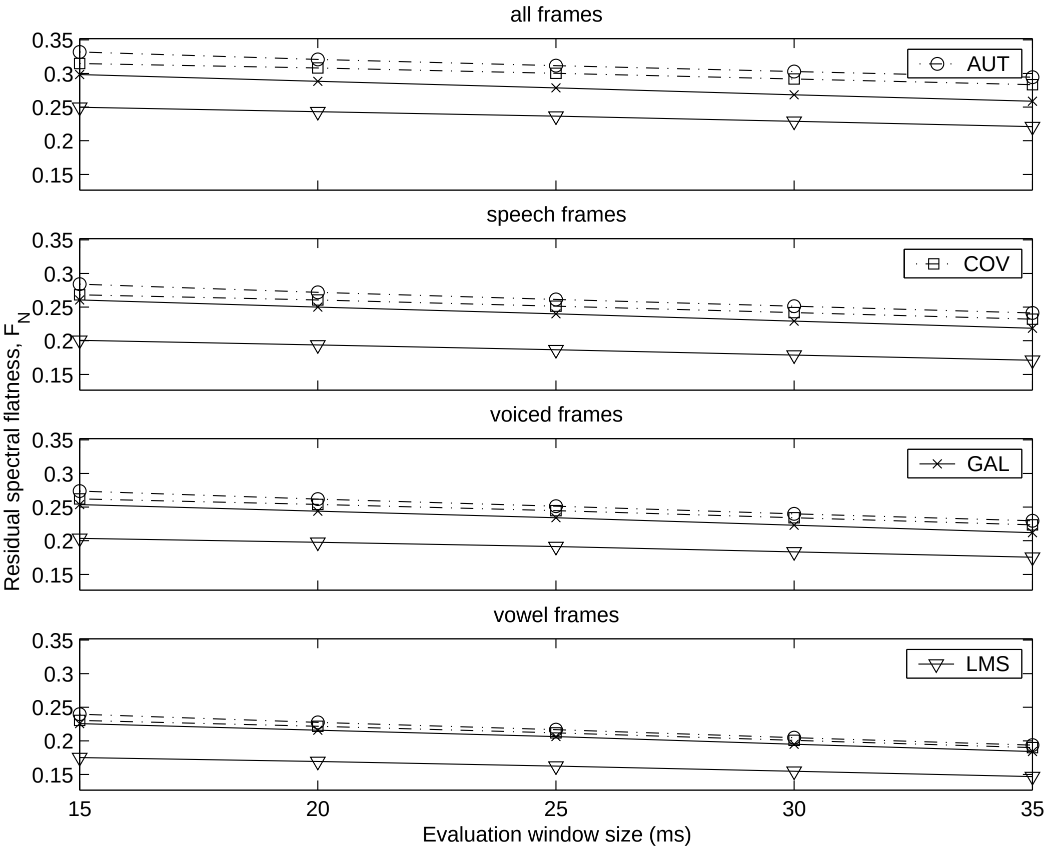

Figure 7.13: Best achieved average F_N, as a function of the evaluation frame length, for four phonetic categories.

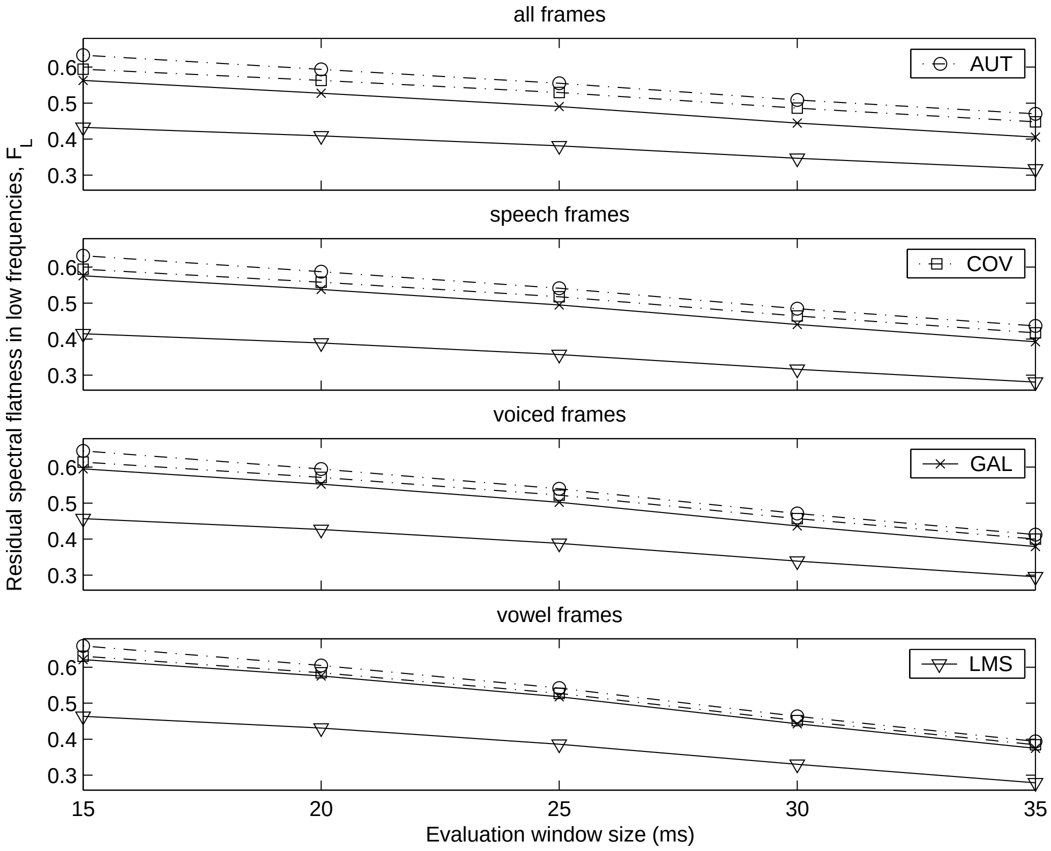

Figure 7.14: Best achieved average F_L, as a function of the evaluation frame length, for four phonetic categories.

Chapter 8

Conclusions

The theoretical part of this thesis offered a treatment of the basic aspects of linear prediction (LP) and warped linear prediction (WLP) especially in the scope of speech spectrum analysis. Objective model quality evaluation measures were also introduced, some of them devised specifically for the purposes of this thesis, and their properties explained.

The different aspects portrayed by the different model evaluation measures were shown in practice by comparing a standard Bark-scale WLP technique with an ordinary LP technique. As an important result, objective criteria were used to select the model order of Bark-WLP so that the average model quality at the relevant formant frequencies corresponds to conventional LP. Because the used criteria were maximally favorable to WLP, prediction order 10 can be regarded as a lower bound for accurately modeling a 22 kHz speech signal using Bark-WLP.

Next, this thesis discussed adjusting the temporal behavior of four WLP techniques: autocorrelation method, covariance method, gradient adaptive lattice, and least-mean-squares. A method was introduced for tuning the temporal resolution of a speech analysis method and verified on the block estimation methods (autocorrelation and covariance); as expected, the chosen WLP analysis frame size had a clear correspondence with the predefined *evaluation frame* size. Then the method was applied to determining the optimal adaptation parameters in the two adaptive estimation methods (GAL and LMS). This is another important result, since as was illustrated, poor selection of the adaptation parameter may lead to very poor results and the behavior of the adaptive methods can be quite sensitive to the value of this parameter.

Finally the four methods, with properly adjusted prediction order and temporal resolution, were compared with each other in terms of relative performance in different phonetic contexts. Block estimation methods are more accurate than the adaptive methods. Without appropriate signal transforms or other modifications, LMS is not the best possible method to use in speech analysis as could have been expected. The allpass transform does not help the situation. Because of its simplicity, LMS may still be an useful starting point for developing more suitable methods. By optimizing the adaptation rate with respect to the average performance, as was done here, GAL produces models whose quality is comparable to the block methods especially in relatively stationary regions in speech such as vowels. However, if good time resolution with e.g. the occlusions of stop consonants is essential, one should consider either a faster-adapting method or some way to automatically adjust the adaptation rate. Nevertheless, each method may be useful in certain applications depending on the requirements on e.g. spectral resolution, time resolution, filter stability, and computational effectiveness.

Bibliography

[1] S. Chandra and W.C. Lin. Experimental comparison between stationary and nonstationary formulations of linear prediction applied to voiced speech analysis. *IEEE Trans. Acoustics, Speech, and Signal Proc.*, ASSP-22:403–415, 1974.

[2] J.R. Deller, J.G. Proakis, and J.H.L. Hansen. *Discrete-Time Processing of Speech Signals*. Macmillan Publishing Company, 1993.

[3] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. John Wiley and Sons Inc, 2nd edition, 2001.

[4] G. Fant. *Acoustic Theory of Speech Production*. Mouton, The Hague, 1960.

[5] L.J. Griffiths. A continuously-adaptive filter implemented as a lattice structure. In *Proc. Int. Conf. Acoust. Speech, and Signal Proc.*, pages 683–686, Hartford, USA, May 1977. IEEE.

[6] J.D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.

[7] M.O. Hayes. *Statistical Digital Signal Processing and Modeling*. John Wiley and Sons Inc., 1996.

[8] S. Haykin. *Adaptive Filter Theory*. Prentice-Hall, 1996.

[9] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *J.Acoust.Soc.Am.*, 87(4):1738–1752, 1990.

[10] A. Härmä. Linear predictive coding with modified filter structures. *IEEE Trans. Speech and Audio Proc.*, 9(8):769–777, November 2001.

[11] A. Härmä, M. Karjalainen, L. Savioja, V. Välimäki, U.K. Laine, and J. Huopaniemi. Frequency-warped signal processing for audio applications. *J. Audio Eng. Soc.*, 48(11):1011–1029, November 2000.

[12] A. Härmä and U.K. Laine. A comparison of warped and conventional linear predictive coding. *IEEE Trans. Speech Audio Processing*, 9(5):579–588, July 2001.

[13] A. Härmä, U.K. Laine, and M. Karjalainen. Backward adaptive warped lattice for wideband stereo coding. In *Signal Processing IX: Theories and Applications, EUSIPCO'98*, pages 729–732, Rhodes, Greece, September 1998. EURASIP.

[14] The MathWorks Inc. The mathworks - matlab, November 2003. http://www.mathworks.com/products/matlab/.

[15] N.S. Jayant and P. Noll. *Digital Coding of Waveforms*. Prentice-Hall, 1984.

[16] M. Karjalainen. Kommunikaatioakustiikka. Technical report, Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, 1999.

[17] M. Karjalainen. Auditory interpretation and application of warped linear prediction. In *Proceedings of Consistent & Reliable Acoustic Cues for Sound Analysis (CRAC),Aalborg, Denmark,* September 2 2001.

[18] M. Karjalainen. S-89.300 Ääniteknologian perusteet, 2003, lecture slides, February 2004. http://www.acoustics.hut.fi/teaching/S-89.300/lectures/.

[19] M. Karjalainen, T. Altosaar, and M. Vainio. Speech synthesis using warped linear prediction and neural networks. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '98*, volume 2, pages 877–880, May 1998.

[20] E. Krüger and H.W. Strube. Linear prediction on a warped frequency scale. *IEEE Trans.Acoustics,Speech,and Signal Proc.*, 36(9):1529–1531, September 1988.

[21] U.K. Laine, M. Karjalainen, and T. Altosaar. Warped linear prediction (WLP) in speech and audio processing. In *Proc. ICASSP-94*, volume iii, pages 349–352, 1994.

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

[23] J. Makhoul. Stable and efficient lattice methods for linear prediction. *IEEE Trans. Acoust., Speech, and Signal Proc.*, ASSP-25(5):423–428, October 1977.

[24] J.D. Markel and A.H. Gray. *Linear Prediction of Speech*. Communication and Cybernetics 12. Springer-Verlag, 1976.

[25] J.S. Milton and J.C. Arnold. *Introduction to probability and statistics : principles and applications for engineering and the computing sciences*. McGraw-Hill, 3rd edition, 1995.

[26] S.S. Narayan, A.M. Peterson, and M.J. Narasimha. Transform domain LMS algorithm. *IEEE Trans. Acoustics, Speech, and Signal Proc.*, ASSP-31(3):609–615, June 1983.

[27] D. O'Shaughnessy. *Speech Communications: Human and Machine*. IEEE Press, 2nd edition, 2000.

[28] L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, 1993.

[29] L.R. Rabiner and R.W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

[30] J.O. Smith and J.S. Abel. The bark bilinear transform. In *Proc. IEEE ASSP Workshop, New Paltz*, 1995.

[31] J.O. Smith and J.S. Abel. Bark and ERB bilinear transform. *IEEE Trans.Speech and Audio Processing*, 7(6):697–708, November 1999.

[32] H.W. Strube. Linear prediction on a warped frequency scale. *J.Acoust.Soc.Am.*, 68(4):1071–1076, October 1980.

[33] M. Ursin. Triphone clustering in finnish continuous speech recognition. Master's thesis, Helsinki University of Technology, 2002.

[34] M. Ursin and U.K. Laine. Comparison of frequency warped front-ends for an HMM based spoken digit recognizer. Unpublished. Subject of a poster presentation for student forum of ICASSP 2001, Salt Lake City, Utah, USA, 2001.

[35] B. Widrow and M.E. Hoff Jr. Adaptive switching circuits. In *IRE WESCON Conv. Rec.*, pages 96–104, 1960.

[36] B. Widrow, J.M. McCool, M.G. Larimore, and C.R. Johnson Jr. Stationary and nonstationary learning characteristics of the LMS adaptive filter. *Proceedings of the IEEE*, 64(8):1151–1162, August 1976.

[37] K. Wiik. *Fonetiikan perusteet*. WSOY, 2nd edition, 1998.