UNIVERSITÄT KALRSRUHE (TH)
INSTITUT FÜR NACHRICHTENTECHNIK

CARNEGIE MELLON UNIVERSITY
INTERACTIVE SYSTEMS LABS

Minimum Variance Distortionless Response Spectral Estimation and Subtraction for Robust Speech Recognition

Diplomarbeit von
Matthias Wölfel

Hauptreferent:	Prof. Kristian Kroschel
Betreuer:	Prof. Alex Waibel, Dr. John McDonough

Beginn:	15.7.2002
Abgabe:	15.1.2003

Matthias Wölfel
Wilhelm-Roether-Str.12
76307 Karlsbad
Germany

Matthias@Wolfel.de

Declarations

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Diplomarbeit selbständig und ohne unzulässige fremde Hilfe angefertigt habe. Die verwendeten Literaturquellen sind im Literaturverzeichnis vollständig aufgeführt.

Karlsruhe, den 14. 1. 2002

Matthias Wölfel

Permission is herewith granted to the Universität Karlsruhe (TH), Germany and Carnegie Mellon University, PA, USA to circulate and to have copies for non-commercial purposes.

Karlsruhe, 1/14/2002

Matthias Wölfel

Abstract

Automatic speech recognition systems that operate in the "real world" are often confronted with acoustic conditions that do not resemble those seen in training data. If the training data is clean, noise that is present at recognition time will result in a mismatch between training and test conditions, and thus lead to a degradation in recognition performance.

This thesis investigates the effectiveness of *spectral envelopes* derived by the *minimum variance distortionless response* (MVDR), both with and without *spectral subtraction*, to reduce the negative impact of such additive noise.

In addition, a novel scaling for the MVDR envelope is proposed, as is the performance of Mel-warping before the calculation of the MVDR.

These investigations have confirmed the superiority of the MVDR to the linear prediction, but the superiority of the MVDR to the Fourier transform could not be confirmed. The methods of spectral subtraction, scaling and pre-warping in combination with MVDR did, however, provide significant improvements in recognition performance as compared with the Fourier transform in most cases of interest.

Zusammenfassung

Automatische Spracherkennungssysteme, werden oft in der "realen Welt" mit einer akustischen Umgebung konfrontiert, die nicht immer der trainierten entspricht. Wurde zum Beispiel mit klaren Sprachdaten trainiert, führen Geräusche während der Erkennung zu einer Abweichung des vorliegenden zum gelernten Muster. Dies kann zu einer fehlerhaften Erkennung und somit zum Verlust an Wortakkuratheit führen.

In der hier vorliegenden Arbeit wird untersucht inwieweit *Spektrale Einhüllende*, berechnet über die *Minimum Variance Distortionless Response* (MVDR), mit und ohne Kombination von *Spektraler Subtraktion* in der Lage sind den negativen Effekt additiver Geräusche zu mindern.

Über die Aufgabenstellung der Diplomarbeit hinausgehend wurden eine neuartige Skalierung der über die MVDR berechneten Einhüllenden vorgeschlagen sowie die Verlegung des Mel-Warpings vor die Berechnung der MVDR Einhüllenden.

Die Untersuchungen konnten die prinzipielle Überlegenheit der MVDR gegenüber der Linearen Prediktion bestätigen, nicht aber gegenüber der Fouriertransformation. Die hier zum ersten Mal untersuchten Methoden, Spektraler Subtraktion, Skalierung und Pre-Warping in Kombination mit MVDR brachten eine leichte Verbesserung der Wortakkuratheit in den meisten hier untersuchten Fällen gegenüber der Fouriertransformation.

Acknowledgements

I would like to express my gratitude to my supervisor Dr. John McDonough for constant support during the research conducted for this thesis. I would also like to thank Prof. Dr. Kristian Kroschel and Prof. Dr. Alex Waibel, the former for the opportunity and encouragement to write a thesis outside the electrical engineering department, the latter for expressing strong interest in this work and for giving me the chance to gain experience and a new perspective through my time spent at the Carnegie Mellon University.

Furthermore, I would like to thank the following:
The speech recognition group at Carnegie Mellon under the guidance of Prof. Dr. Richard Stern; my officemates Qin Jin, Yue Pan and Bing Zhao; Sebastian Stücker, Hua Yu and all the other wonderful people at the Interactive Systems Labs.

Last but not least, I am grateful to my parents, Doris and Helmut Wölfel, for their endless support, trust and *love*. Without them this work would literally never have come into being.

Table of Contents

List of Tables

List of Figures

1 Introduction

In an ideal environment, where there is a single speaker wearing a head-mounted, close-talking microphone, *automatic speech recognition* (ASR) achieves recognition rates up to 99%, which are eminently usable. When confronted with "real world" conditions, however, recognition rates drop significantly, because many of acoustic conditions seen in actual use do not resemble those present in the training data. Significant causes of mismatch between training and test conditions are ambient noise, reverberation and speaker variation. A similar mismatch will occur if the noise at training differs from the noise occurring at recognition. Because such mismatches degrade recognition performance, additional techniques, applied either before recognition or as an integral component thereof, are needed to mitigate their harmful effects. Loosely speaking, we can identify three different approaches for increasing robustness:

1. **Feature-based approach**

   This approach attempts to increase the robustness of the features used for ASR, which can be achieved in two different ways:

   - **Robust feature selection**

     Selection of features that are relatively insensitive to the unwanted variations, while eliminating features which are sensitive and thus contain significant distortions which may result in an errorful decision.

   - **Speech or feature enhancement**

     In this approach the speech signal is manipulated in a way to enhanced quality for the ASR; e.g.. spectral subtraction.

   Essentially the rest of the system is the same as for the recognition of clean speech.

2. **Model-based approach**

   In a model-based approach, all information concerning various kinds of variability is retained after the acoustic pre-processing. One of the model-based approaches for robustness of ASR is to separately model the different parts of speech by different parts of the model structures. For example, separate sets of *hidden Markov models* (HMM)s parameters (e.g. the mean and variance of the Gaussians) are defined for the useful information and the distortion part, respectively.

3. **Microphone-array approach**
   In this case, the additional structure that tries to separate the speech and distorting signals is physically located outside the recognizer. The geometrical relations between the different sound sources will result in different signals arriving at the microphones in the array which makes it possible to focus on one particular direction and thus suppress noises from other directions.

1.1 Review of Prior Work

In an investigation by de Wet, Cranen, de Veth, and Boves it was found that *Mel-frequency cepstral coefficients* (MFCC) derived by the *spectral envelope* (SE) based on *linear prediction* (LP) could lead to features which are inherently more robust to at least some kind of background noise than their counterparts derived from the Fourier transform [Wet00][Wet01]. Although LP is popular in speech recognition, it tends to overestimate and overemphasize the spectral peaks in medium- and high-pitched voiced speech [Mur00, Kab00], and provides a resolution far beyond the Fourier transform. These drawbacks result in a loss of recognition accuracy in clean speech.

Murthi and Rao introduced a SE based on the *minimum variance distortionless response* (MVDR), which can be easily obtained from the *LP coefficients* (LPC)s. Unlike LP, the MVDR provides an elegant envelope representation of the spectrum for both medium- and high-pitched voices [Mur97]. Furthermore, combining the MVDR with a smoothing technique for reducing the variance in the features can improve the *word error rate* (WER) in *robust speech recognition* (RSR) [Dha01].

An all-pole model approximates spectra equally well at all frequency bands. Usually the spectrum is warped after all-pole analysis, which does not lead to an improvement of the frequency resolution of the envelope. To improve the frequency resolution in low frequencies, Strube [Str80] proposed a method based on a linear transform to apply warping before the all-pole analysis. Applied to LP using the Mel-frequency as a warping factor and thus dubbed Mel-LP, this approach could provide a significant improvement in recognition accuracy over LP and a slightly higher recognition accuracy for male speakers over Mel-frequency cepstral coefficients [Mat01].

Boll [Bol79] proposed the technique of *spectral subtraction* (SS), where an estimate of the noise spectrum is subtracted from the spectrum of the noisy signal. It has been shown that this method can successfully increase the *signal-to-noise ratio* (SNR) of speech signals corrupted by additive noise [Ars95]. Unfortunately, the remaining signal tends to consist of short duration random tones, known as *musical tones/noise*. The appearance of musical tones in a speech signal can be very objectionable to human ears and limit the gain in word accuracy of a ASR system achieved by the increased SNR. To suppress the appearance of musical tones, a broad variety of techniques were suggested [Cap94][Kot01].

1.2 Contributions of this Work

This thesis aims to reduce the negative impact of noise on recognition performance following the feature-based approach, or to be more specific, by enhancing the speech features. Based on the MVDR spectrum additional methods were suggested, implemented into the *Janus Recognition Toolkit*[1] (JRTK) and investigated:

1. **Spectral subtraction based on the spectral envelope to suppress musical tones**
   Most approached to overcome musical tones are post-processing steps which try to limit the effects of musical tones after their appearance. Here we propose and investigate spectral subtraction (SS) based on the SE instead of the Fourier transform, which is similar to the smoothing of the spectrum already suggested by Boll [Bol79]. Similar to the smoothing, the SE reduces the variance of the spectral estimate and therefore successfully prevents musical tones from occurring. It also overcomes the drawback of spectral smoothing which decreases the spectral accuracy resulting in smeared *formants* which are very important in speech recognition.

[1] For a detailed description see Chapter 7.2

2. **Introduction of a novel scaling technique of the MVDR envelope**
   The MVDR envelope is able to overcome the main problems of the LP envelope in medium and high pitched voices, but a high variance of the amplitude remains. To reduce this variance we propose a novel scaling technique which adjusts the maximum of the envelope to the maximum of the Fourier spectrum, which in average is less distorted than the averaged energy.

3. **Adaptation of pre-warping to the MVDR approach**
   Mel-LP do not provide a good SE for medium- and high-pitched voiced speech. High order MVDR all-pole models have been shown to be superior to LP all-pole models for medium and high pitched voiced speech, and therefore an adaptation of high order MVDR all-pole models to "*Mel-MVDR*" could provide a spectrum envelope which models medium and high pitched voiced speech very well in the Mel-spectra. The Mel-MVDR followed by a filterbank could also lead to an improved frequency resolution in low frequencies over the MVDR followed by a Mel-filterbank leading to a better word error rate.

To measure and compare recognition performances of the proposed techniques, the JRTK using different speech recognition pre-processing approaches was trained on clean speech and tested in different acoustic conditions. Adverse acoustic conditions were simulated by "artificially" adding noise (white noise at different SNRs and noise recorded from robots) and through recordings in a meeting room. A second data set containing telephone speech was also investigated.

It should be kept in mind that observations with "artificial" disturbance may not always generalize to "real world" applications; clean signals with "artificially" added noise are by no means exact representations of "real world" noise conditions. For instance, they do not capture the way in which people tend to change their rate and manner of speaking if the acoustic condition gets noisy (Lombard effect) [Lom11] entailing a higher pitch, slightly different formants and a different coloring of the spectrum. Nevertheless, these simulations are widely used for experimental purposes, because they provide a framework within recognition performance in clean and noisy acoustic conditions may easily be compared. Such a framework also provides the possibility to measure the impact of additive noise on the statistical properties of the data at acoustic feature level.

1.3 Organization of this Work

We now outline the balance of this thesis. Chapters 2–4 review briefly well known material to lay the ground work for the development which follows. Chapter 2 reviews quality factors in speech recognition which allow system performances to be measurable. Chapter 3 reviews the components of a widely used model of acoustic pre-processing in ASR and reviews and discusses how the aspects of the human auditory system are or could be implemented in the pre-processing. Chapter 4 reviews the concepts of LP and its interpretation in the frequency-domain. Chapter 5 reviews the basic ideas of the MVDR and offers a fast way for its computation. Furthermore, novel MVDR scaling and warping techniques are proposed and their effects are discussed. Chapter 6 reviews spectral estimation and subtraction and introduces SS based on the envelope and discusses why it should perform superior. In Chapter 7, speech recognition experiments are conducted to measure the recognition performance of the novelties which are so far only discussed theoretically. Finally, Chapter 8 summarizes the work and suggests ways in which the MVDR approach might lead to further improvement in word accuracy in speech recognition and suggest the use of the MVDR in other applications.

2 Quality Factors

To quantify the effectiveness of the robustness measures proposed in this work, quality factors are needed which correspond well with the performance of an ASR system and/or the quality of the processed speech. Such measures fall into two primary categories:

1. **Subjective Quality Measures**

   This category of quality measures is based on the opinion of a group of test subjects and can be classified as:

   - **Listening Test** where a group of listeners judge the quality and/or intelligibility of the speech. Here, a consistent listening environment is required since the perceived distortion can vary with the playback volume and type of listening instrument; e.g., headphones, speakers.
   - **Understandability** where a group of persons render opinions about the understandability of the output of an ASR system. A subjective judgment is necessary in this case, because the meaning of a sentence may be preserved even if one or more words are incorrectly recognized or deleted.

   In general, these measures are time-consuming and costly.

2. **Objective Quality Measures**

   This category of quality measures can be evaluated automatically from the speech signal, its spectrum or some parameters obtained therefrom. Since they do not require listening tests, these measures can give an immediate estimate of the perceptual quality of an algorithm. In addition, they can serve as a mathematically tractable criterion to minimize. The two main factors in selecting an objective distortion measure are performance and complexity. The performance of an objective distortion measure can be established by its correlation with a subjective distortion measure of the same features (quality or intelligibility). Objective quality measures can be broadly classified into four categories:

- **Time-domain distortion measures** are most useful for waveform coders which attempt to reproduce the original speech waveform. The most frequently encountered measures of this type are the several forms of signal-to-noise ratio.
- **Frequency-domain distortion measures** are used to determine the performance of the magnitude spectrum; e.g., log spectral distortion.
- **Word Error Rate & Word Accuracy** are the most important measures of an ASR system, because they directly state the recognition performance.
- **Perceptual-domain distortion measures** are based on human auditory models. They transform the signal into a perceptually relevant domain and take advantage of psychoacoustic masking effects.

Performance analysis of distortion measures are given in [Dim89][Dim95][Qua88].

2.1 Time-Domain Distortion Measures: Signal-to-Noise Ratio

The *signal-to-noise ratio* (SNR) represents an average error over time or frequency for a processed signal and is defined as the ratio between the power of the signal to the power of the noise. Usually the SNR is defined in *decibels* (dB) as:

$$SNR = 10\log_{10}\frac{E_S}{E_N} = \frac{\sum_{n=-\infty}^{\infty} s^2[n]}{\sum_{n=-\infty}^{\infty}(s[n]-\hat{s}[n])^2} \qquad (2.1.1)$$

where $\hat{s}[n]$ denotes the estimate of the original speech sample $s[n]$. The principal benefit of the SNR quality measure is its mathematical simplicity. The fact that the SNR is not particularly well related to any subjective attribute of speech quality and that it weights all time domain errors in the speech waveform equally, makes it a poor measure for a broad range of speech distortions. Furthermore, in the case of speech recognition, an improvement in the SNR does not necessarily increase the *word accuracy* (WA).

2.2 Frequency-Domain Distortion Measures

Generally speaking, the spectral distortion should measure the discrepancies between the original signal and its estimated version, which will unavoidably contain distortions that might lead to errors in phonetic classification [Rab93]. The disparities between the original signal and its estimate may include the following:

- Significant differences in the center frequencies of *resonances* or *formants*.
- Alteration of the formant bandwidths due either to the distortion or the estimation process.

To measure spectral distortion, a function $d(f, \hat{f})$ of two spectral densities, the true spectrum f and it's counterpart $\hat{f}$, is defined with following properties:

- Non-negativity: $d(f, \hat{f}) \geq 0$
- $f = \hat{f} \Leftrightarrow d(f, \hat{f}) = 0$
- Symmetry: $d(f, \hat{f}) = d(\hat{f}, f)$
- Satisfaction of the triangular inequality: $d(f, g) \leq d(f, h) + d(h, g)$

The final performance measure is then the long term average of a given distortion measure, expressed as follows:

$$D = \lim_{M \to \infty} \frac{1}{M} \sum_{m=1}^{M} d(f_m, \hat{f}_m) \tag{2.2.1}$$

One of the important spectral distortion measures which is also used throughout this thesis is discussed below. Further examples of spectral distortion measures can be found in [Bat98].

2.2.1 Logarithmic Spectral Distortion

The *logarithmic spectral distortion* (LSD), for a given frame, is defined as the root mean square difference between the original *logarithmic power spectrum* (LPS) S and the estimated LPS $\hat{S}$. Mathematically, the L_p norm-based logarithmic spectral distance measure is defined as

$$d_{LSD}^{p} = \frac{2}{F} \int_{0}^{F/2} |\log_{10} S(\omega) - \log_{10} \hat{S}(\omega)|^{p} d\omega \tag{2.2.2}$$

where F denotes the sampling frequency. As usual, for the discrete case we can replace the integration with a summation. For the balance of this thesis, the LSD will be defined as the second $(p = 2)$ norm-based LSD, such that:

$$\text{LSD} = \sqrt{\frac{1}{M} \sum_{m=1}^{M} [\log_{10} S(m) - \log_{10} \hat{S}(m)]^{2}} \tag{2.2.3}$$

Paliwal and Atal [Pal93] have suggested that the average spectral distortion alone is not adequate to measure perceived quality. They introduced the notion of spectral outliers which represent the fraction of frames with large spectral distortions.

2.3 Word Error Rate & Word Accuracy

The most often cited performance measure for an ASR system is *word error rate* (WER), which is typically expressed as the percentage:

$$\text{WER} = \frac{N_{err}}{N} = \frac{N_{sub} + N_{del} + N_{ins}}{N} \tag{2.3.1}$$

or the *word accuracy* (WA), which is given by:

$$\text{WA} = 100\% - \text{WER} = \frac{N - N_{\text{sub}} - N_{\text{del}} - N_{\text{ins}}}{N} \tag{2.3.2}$$

In the above, N denotes the number of tokens, N_{sub} the number of substitution errors, N_{del} the number of deletion errors and N_{ins} the number insertion errors.[1]

2.3.1 Relative Error Reduction

When comparing different acoustic modeling algorithms, it is not useful to use absolute WERs, because a reduction of 1% in WER from, e.g., 3% to 2% WER is more significant than a reduction from, e.g., 30% to 29% . Therefore a measure is more useful which is able

[1] In the case of non-continuous speech recognition N_{del} and N_{ins} can't appear and thus are set to zero.

to state the significance of the error reduction. This can be achieved by the *relative error reduction* (RER):

$$\mathrm{RER} = \frac{\mathrm{WER}_B - \mathrm{WER}_A}{\mathrm{WER}_B} \tag{2.3.3}$$

For the given example, the RER would be 33.3% (WER reduced from 3% to 2%) compared to 3.3% (WER reduced from 30% to 29 %).

Empirically, you need to have a test set containing more than 8000 words from at least 10 different speakers to reliably estimate the WER and RER [Hua01].

2.4 Computation Cost & Memory Usage

Among the WA of an ASR system, factors like *computation cost* and *memory usage* play a major role. Usually for the computation cost the *real time factor* (RTF) is used; e.g., an RTF of 2.0 means that the computer needed twice as long to recognize a recording as the speaker took to say it. As my investigations are only addressing a small part of an ASR system and the RTF is strongly dependent on the used system, we prefer to compare the calculation cost between different implementations. Memory usage should not addressed, because in my investigations it plays no mayor role as it is always very low.

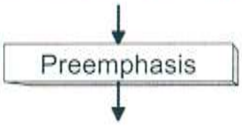
3 Acoustic Pre-Processing

3.1 Acoustic Pre-Processing for Speech Recognition

In this chapter we introduce the acoustic pre-processing in a speech recognizer, also termed *speech recognition frontend* because it refers to the first stage of an ASR system. Its task is to transform the acoustic input signal to a sequence of *acoustic feature vectors* preserving all the perceptually important information for phonetic distinctions, while being insensitive to phonetically irrelevant variations.

Over the years many different speech recognition frontends have been developed. The variety of frontends are distinguished by the extent to which they incorporate information about the human auditory processing and perception. Figure 3.3 shows a widely used ASR frontend, which is also the basic approach in the JRTK used throughout the evaluations presented in this thesis.

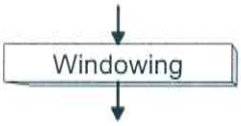
The speech waveform is commonly sampled at 16 kHz, 16-bit A/D precision which is sufficient for the speech bandwidth of 8 kHz. Reducing bandwidth generally decreases the word accuracy; e.g., a down sampling from 16 kHz to 8 kHz (typical for telephone speech) results in an increase of relative word error by 20% [Hua01].



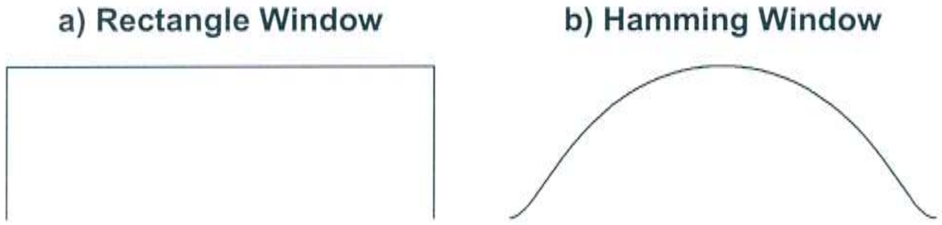
To compensate for the unusual sensitivity of human hearing across frequency equally-loudness pre-emphasis a first-order high pass filter

$$H(z) = 1 - \alpha \cdot z^{-1} \tag{3.1.1}$$

is used [Mil02].



Speech is a so-called *quasi-stationary* signal, which implies that the vocal tract shape, and thus its transfer function, remain nearly fixed over short time intervals of 5-25 ms duration. The signal processing performed by a typical ASR frontend assumes a stationary signal; hence it is necessary to split a given utterance into short segments. The length of each segment must be short enough to give the required time resolution, which involves a tradeoff with adequate frequency resolution. In addition, during voiced speech the signal must be long enough to be insensitive to exact positioning relative to the glottal cycle. The advantages of a long segment is that it smooths out some of the temporal variations of unvoiced speech while it blurs rapid events such as the releases of stop consonants.


a) Rectangle Window

b) Hamming Window

Figure 3.1: Rectangle and Hamming window

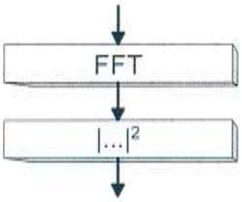
To apply the segmentation, the total speech signal is multiplied every 10 ms (frame shift) by an analysis window of a fixed duration between 16 and 25 ms (frame size). Choosing the right window shape is very important, as this shape determines the resolution of the speech segment in the frequency domain. The simplest analysis window is a *rectangular window*:

$$w(n) = \begin{cases} 1 & 0 \leq N_w - 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.1.2}$$

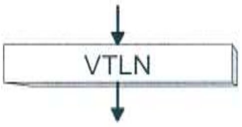
Because of its abrupt discontinuity at the edges, the rectangular window introduces spurious high-frequency components in the frequency domain. To reduce the discontinuities at the edges of the selected regions, a window without abrupt discontinuities in the time domain should be used. One window of this type commonly used in speech processing is the *Hamming window*:

$$w(n) = \begin{cases} 0.54 - 0.46\cos(\frac{2\pi n}{N_w - 1}) & 0 \leq N_w - 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.1.3}$$

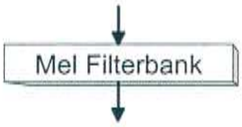
Descriptions of other types of tapered windows (e.g., Hanning, Blackman, Kaiser, Bartlett) can be found in any digital signal processing book [Kam98, Opp89].



To derive the PS, first the *fast Fourier transform* (FFT) is computed, then it's absolute value is squared.



Due to gross differences in the length of the vocal tract, which is strongly correlated with a person's height, the locations of the spectral resonances or formants seen in voiced speech vary widely across speakers. To reduce inter-speaker variability a *vocal tract length normalization* (VTLN) may be applied shift the formants for a given speaker back to their nominal locations. This shift, also called a *warping factor*, is estimated for each speaker by computing the likelihood of the training data for feature sets obtained with different shifts. This simple normalization can provide a relative reduction in WER of as much as 10% [Hua01].



To immitate the frequency dependent spectral resolution of the human ear, the PS is warped according to the *Mel-scale* by the following transformation

$$f_{\text{Mel}} \triangleq 2595 \log_{10}\left(1 + \frac{f_{\text{Hz}}}{700}\right) \tag{3.1.4}$$

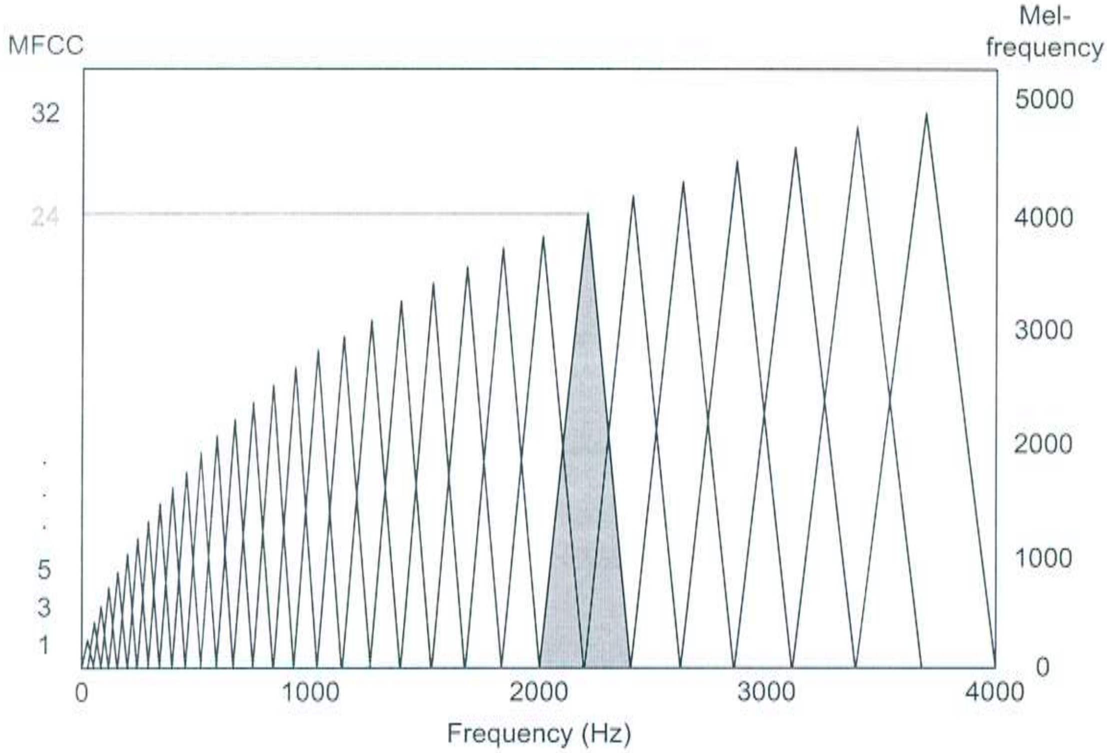
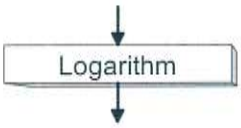


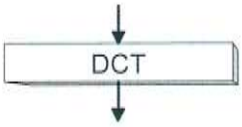
Figure 3.2: Critical band filters for the Mel-frequency/-cepstrum

To reduce the number of features, a filterbank of uniformly half overlapping triangular shaped filters could be used.

A smarter solution combines Mel-warping and feature reduction by placing the triangular filters non-uniformly at the unwarped spectrum, as in Figure 3.2, and thereby implicitly incorporate Mel-frequency scaling [Sha87]. This *Mel filterbank* will be used throughout the remainder of this thesis, unless the contrary is specifically stated.



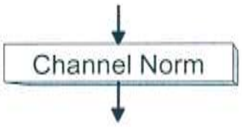
This processing stage models the non-linear relation between the intensity of sound and its perceived loudness.



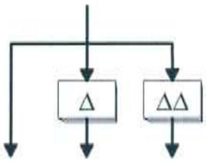
A further reduction of recognition errors may be obtained through a transformation of the feature vectors into a new space that is less sensitive to environmental noise, channel distortion, and speaker variations. One useful transformation of this type is the *discrete cosine transformation* (DCT) which transforms the feature vectors into the *cepstral space* where the *cepstral coefficient* are calculated as

$$c_m = \sqrt{\frac{2}{N}} \sum_{n=1}^{N} A_n \cos\left(\frac{\pi n(m-0.5)}{N}\right) \qquad \forall\, 0 \leq m \leq N-1 \tag{3.1.5}$$

resulting in the *Mel-frequency cepstral coefficient* (MFCC) vector. The cepstral sequence is truncated to 12-15 components to smooth the spectrum and to minimize the influence of the pitch which is irrelevant for the speech recognition process.



The channel is normalized by a subtraction of the mean for each cepstral component. This step reduces the RER by 5% while increasing the robustness of the system to different environments [Lee96]



Because temporal changes in the spectra play an important role in human speech perception, the static cepstral coefficients are typically augmented by first and second order delta coefficients, which measure the change in the static coefficients over time. This leads to an RER of 20% [Hua01].

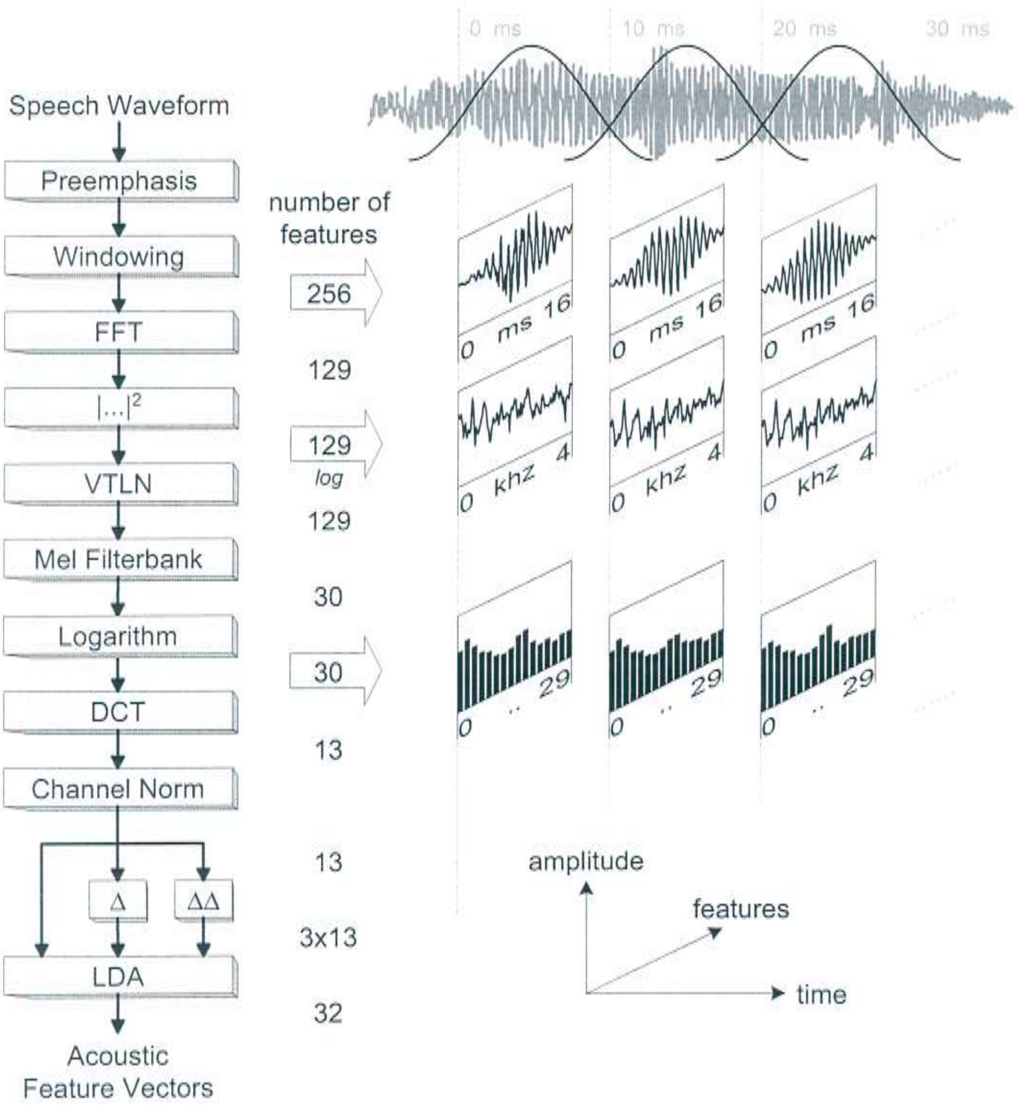
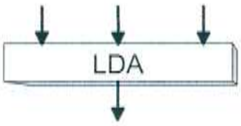
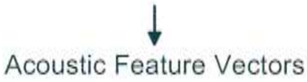


Figure 3.3: *Typical* frontend of a speech recognition system



To further reduce the dimension of the feature vector, a dimension reduction technique to map the feature vector into a more effective representation may be used. A simple criterion is to use *within-class* and *between-class* scatter matrices to formulate criteria of class separability, this is also referred to *linear discriminant analysis* (LDA). For a detailed description of this technique, see [Hua01].



The acoustic features are then used for higher level processing; see [Rab93, Hol01].

Further details of the acoustic pre-processing for speech recognition can be found in [Hol01, Mol01].

3.2 Aspects of the Human Auditory System

It is widely known that in speech recognition an adaptation of the aspects of the human auditory system can reduce calculation costs and increase word accuracy [Her90]. Therefore, in this chapter, we want to review several aspects of the human auditory system and discuss how the same can be applied to an ASR system.

- **Phase insensitivity**
  The phase components of a speech signal play a negligible role in speech perception, with weak constraints on the degree and type of allowable phase variations [Del00]. The human ear is fundamentally phase "deaf" and perceives speech primarily based on the magnitude spectrum.

  This can easily applied in a speech recognition approach by using the absolute of the complex spectrum and justifies the use of a minimum-phase system to represent the possibly non minimum-phase impulse response of the vocal tract.

- **Perception of spectral shape**
  Spectral peaks (corresponding to poles in the system function) are more important to

perception than spectral valleys (corresponding to zeros) [Sai85].

This can be applied by using an all-pole model as we will see in Section 4.1.

- **Frequency masking**

  Every short-time PS has an associated masking threshold. The shape of this masking threshold is similar to the spectral envelope of the signal, and any noise inserted below this threshold is "masked" by the desired signal and thus inaudible.

  This feature may be applied by a spectral envelope.

- **Frequency dependent spectral resolution**

  Spectral information in the human auditory system is processed on a non-uniform frequency scale.

  This can be applied by frequency-warped spectral features; e.g., by the Mel filterbank.

- **Temporal masking**

  Sounds can mask noise up to 20 ms in the past (backward masking) and up to 200 ms in the future (forward masking) given that certain conditions are met regarding the spectral distribution of signal energy [Sha00].

  As far as we know this principle has not been applied to speech recognition yet.

3.3 Speech Production Model

Knowledge of the vocal system and the properties of the resulting speech waveform is essential in designing an approximate model of speech production.

Due to the inherent limitations of the human vocal tract, speech signals are highly redundant and contain a variety of different, speaker dependent speech parameters, e.g., pitch, formants, spectra, phase and vocal tract area function. By removing the irrelevant information, contained in the waveform, a simple model of human speech production is obtained. In the case of ASR, for example, only the formants and the spectra are of interest.

The human speech production process reveals that the generation of each *phoneme*, the basic linguistic unit, is characterized by two basic factors:

- the random noise or impulse train excitation
- the vocal tract shape

In order to model speech production, we must model these two factors. To understand the source characteristics, it is assumed that the source and the vocal tract model are independent [Del93].

Speech consists of pressure waves created by the flow of air through the vocal tract. These pressure waves originate in the lungs as the speaker exhales. The vocal folds in the larynx can open and close quasi-periodically to interrupt this airflow. This results in *voiced speech*, which is characterized by its periodic and tends to have relatively high energy. Vowels are typical examples.

Some consonants like /f/, /s/ (here /·/ denotes a phoneme) on the other hand are examples of the so called *unvoiced speech*. These sounds are noisy in nature due to turbulence created by the flow of air through a narrow constriction in the vocal tract. The positioning of the vocal tract articulators acts as a filter, amplifying certain sound frequencies while attenuating others.

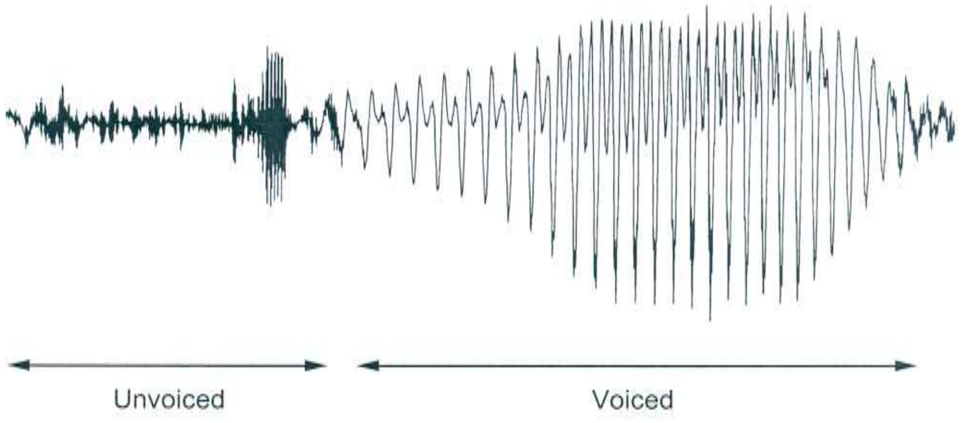


Figure 3.4: A speech segment (time domain) of unvoiced and voiced speech

A time-domain segment of unvoiced and voiced speech is shown in Figure 3.4. A general linear discrete-time system to model this speech production process is shown in Figure 3.5.

In this system, a vocal tract filter $V(z)$ and a lip radiation filter $R(z)$ are excited by a discrete-time excitation signal. The local resonances and anti-resonances are present in the vocal tract filter $V(z)$ which has an overall flat spectral trend. The lips behave as a 1st order high-pass filter and thus the lip radiation filter $R(z)$ grows at 6 dB/octave.

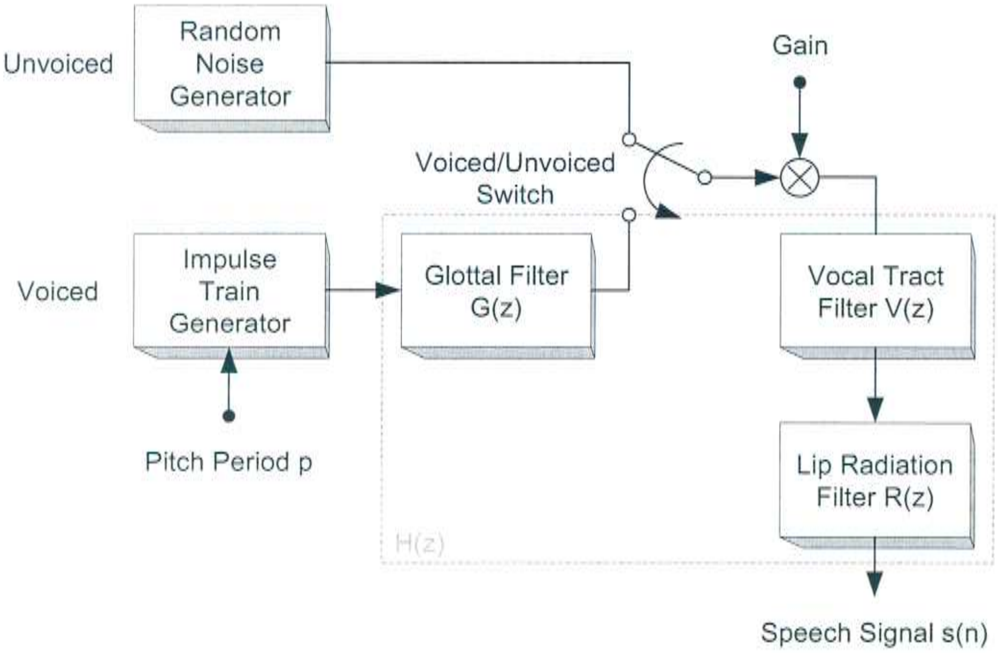


Figure 3.5: Block diagram of the simplified source filter model of speech production

To get the excitation signal for unvoiced speech, a random noise generator with a flat spectrum is typically used. In the case of voiced speech the spectrum is generated by an impulse train with pitch period p and an additional glottal filter $G(z)$. The glottal filter is usually represented by a 2nd order low-pass filter, falling off at 12 dB/octave.

The periodicity of voiced speech gives rise to a spectrum containing harmonics of the fundamental frequency of the vocal fold vibration. A truly periodic sequence, observed over an infinite interval, will have a discrete-line spectrum but voiced sounds are only locally quasi-periodic. The resonances in the PS of voiced speech, known as *formants*, are a product of the shape of the vocal tract. The spectrum for unvoiced speech ranges from flat spectra to those lacking low frequency components. The variability is due to place of constriction in the vocal tract for different unvoiced sounds – the excitation energy is concentrated in different spectral regions. Due to the continuous evolution of the shape of the vocal tract, speech signals are nonstationary. The gradual movement of vocal tract articulators, however, results in speech that is quasi-stationary over short segments of 5-25 ms which allows a splitting of the speech signal in short frame segments of 16-25 ms to perform frequency analysis.

3.4 Spectral Envelope

A *spectral envelope* (SE) is a curve in the amplitude-frequency plane of the signal energy with following desirable properties:

- **Envelope fit**
  The curve of the SE should wrap tightly around the PS, linking the peaks. If it is not possible to link every peak, e.g., when the additive analysis finds a group of peaks close to each other with high energies, then it should find a resonable intermediate path.

- **Robustness**
  The estimation method to derive the envelope has to be applicable to a wide range of signals with very different characteristics, from high pitched harmonic sounds with their wide spaced partials to noisy sounds or mixtures of harmonic and noisy sounds.

- **Smoothness**
  The SE should provide a certain smoothness. This means it must not oscillate too much, but it should give a general idea of the distribution of the signal energy over frequency.

- **Stability**
  The estimation method to derive the envelope should be stable.

- **Locality**
  The SE should be local which states that it should be possible to achieve a local change of the SE, i.e., without affecting the intensity of frequencies further away

from the point of manipulation. Ideally, the representation would fulfill the requirement of orthogonality, where one component of the SE can be changed without affecting the others at all.

- **Speed of synthesis**
  The calculation cost to derive the SE should be as small as possible.

- **Insensitivity to noise**
  The requirement of insensitivity to noise mandates that the representation be resilient to small changes in the data to be represented. Small changes, e.g., in the presence of noise, must not lead to big changes in the representation, but must result in equally small or even smaller (see chapter 3.4.2) changes.

- **Minimal Variance**
  The variance of the envelope of the same phoneme should be as small as possible.

3.4.1 The Advantage of Spectral Envelopes over Smoothed Spectra

Smoothing a spectrum results in three drawbacks which can be overcome by the SE [Gu01]:

- **Limited ability to remove undesired harmonic structures**
  In order to maintain adequate spectral resolution, the standard filter bandwidth is usually in the range of 200Hz-300Hz in the low frequency region. Hence, it is sufficiently broad for typical male speakers, but not broad enough for high pitch (up to 450Hz) female speakers. Consequently, the formant frequencies are biased towards pitch harmonics and their bandwidth is misestimated.

- **The characteristics of the vocal tract is widely believed to be the spectral envelope and not the gross spectrum**
  It is widely agreed in the speech community that it is the SE and not the gross spectrum that represents the shape of the vocal tract [Jel99]. Although the smoothed spectrum is often similar to the SE of unvoiced sounds, the situation is quite different in the case of voiced and transitional sounds. Experiments show that this mismatch substantially increases the spectrum variation within the same utterance. This phenomenon is illustrated in Figure 3.6 which demonstrates that the imaginary

"upper" envelope of the PS sampled at pitch harmonics is nearly unchanged, while the variation of the imaginary "lower" envelope is considerable. The conventional smoothed spectrum representation may be roughly viewed as averaging the "upper" and "lower" SEs. It therefore exhibits much more variation than the "upper" SE alone.

- **High spectral sensitivity to background noise**
  As mentioned in the former paragraph the conventional smoothed spectrum may be roughly viewed as averaging the "upper" and the "lower" SEs. Because regions with low energy are stronger distorted than the regions with high energy (see chapter 5.3) the combination of "both" SE exhibits lower SNR as the "upper" SE alone.

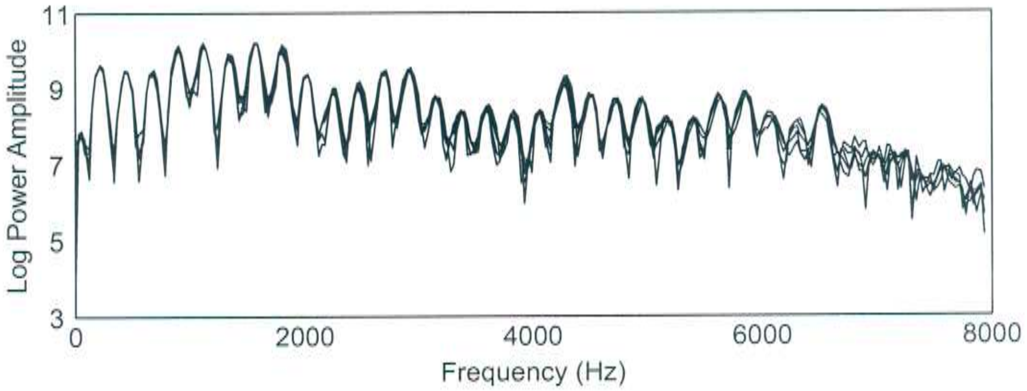


Figure 3.6: Power spectrum of 5 consecutive frames (2ms steps) over a stationary part

3.4.2 Spectral Envelopes in Noisy Environment

Distortion effects due to additive noise in the logarithmic power domain is most evident in the spectral valleys. The spectral peaks, on the other hand, remain relatively unchanged (for an explanation see Section 5.3). The main difference between the SE and Fourier transform is this difference in the description of the spectral valleys:

The SE describes the peaks in the spectrum with high accuracy while the representation of the valleys includes no detailed information about their fine spectral structure. In contrast, non-parametric descriptions of spectra, such as the Fourier transform, describe spectral peaks and valleys in equal detail.

Therefore, if the spectral valleys are filled by additive noise, the spectral fluctuations introduced by the noise should have little effect on SEs while in the Fourier case the spectral fluctuations are described in just as much detail as the spectral peaks, resulting in a higher overall distortion of the spectrum in comparison to the clean spectrum.

4 Linear Prediction Spectral Estimation

The classic method of spectral estimation considers the spectra of short time sections under the implicit assumption that the data values outside the window are zero or repeat the spectra cyclical. This is usually not the case and results in a smeared spectrum and sidelopes [Kay88]. Therefore, to increase the spectral resolution and to suppress sidelopes, an estimation of the values outside the window is needed. *Linear prediction* (LP) is a popular method to achieve this.

The term linear prediction was coined by Wiener in 1966 [Wie66] and applied for speech analysis two years later by Itakura and Saito [Ita68]. Since then it has become one of the most powerful speech analysis and coding methods.

4.1 Linear Predictive Modeling of Speech Signals

Ideally, the output of a prediction filter $H(z)$ in Figure 3.5 should correspond to the physical excitation of the vocal tract that produced the speech segment. Limitations of the model and the error introduced in estimating the model parameters, however, allow only an approximation to the actual excitation signal. The selection of the order p of the LP model is a trade-off between spectral accuracy, computational complexity and, in the case of speech coding, transmission bandwidth. As a general rule, two poles are needed to represent each formant and two to four additional poles are used to approximate spectral nulls (where applicable) and for overall spectral shaping. Based on simple acoustic tube modeling of the vocal tract, the first formant occurs at 500 Hz and the remaining formants occur at roughly 1 kHz intervals (i.e., 1.5 kHz, 2.5 kHz, ...). Therefore, a model order of 8 to 14 is typically used for a sampling frequency of 8 kHz, to model the first three to five formant peaks. Typically this model order increases with higher sampling rates; e.g., at 16 kHz a model order of 20 is common. Conversely, in the case of speech corrupted by noise, studies have shown that the order of the LP all-pole filter should also increase, to model both, speech and noise [Tie80].

To mimic the human auditory system, which has higher resolution in lower frequencies and lower resolution in higher frequencies, a selective LP analysis was developed by Makhoul [Mak73, Sch75]. Its function is to apply the LP analysis on a selected portion of the spectrum rather than uniformly over the entire spectral range. Another approach inspired by the frequency-dependent resolution present in the human auditory system is to pre-warp [Str80, Kar01] the spectrum before linear prediction.

4.1.1 Basic Principles of Linear Prediction Analysis

Referring to Figure 3.5, the combined spectral contribution of the glottal pulse, the vocal tract and the radiation of the lips can be represented by a time varying filter with a steady state system function, as given by a pole-zero model, which is also known as an *autoregressive moving average* or *ARMA model*:

$$H_{\text{pole-zero}}(z) = G \cdot \frac{1+\sum_{j=1}^{M} b_j z^{-j}}{1-\sum_{i=1}^{N} a_i z^{-i}} \tag{4.1.1}$$

Here, poles as well as zeros exist in the transfer function. As previously mentioned, an all-pole model gives a good approximation for a speech signal and provides for increased noise robustness. Thus, we can simplify to an all-pole model, which is also known as *autoregressive* or *AR model*:

$$H_{\text{all-pole}}(z) = \frac{G}{1-\sum_{i=1}^{N} a_i z^{-i}} = \frac{G}{A^{(N)}(z)} \tag{4.1.2}$$

Transforming (4.1.2) into the sampled time domain we get the *LPC difference equation*

$$s(n) = G \cdot x(n) + \sum_{i=1}^{N} a_i s(n-i) \tag{4.1.3}$$

This equation states that the value of the present output $s(n)$ is dependent on the gain G, the present input $x(n)$ and a weighted sum of the past output samples $s(n-i)$, $i = 1, \ldots, N$. Hence, the problem of linear prediction can be stated as the determination of the parameters $a_i,\ i = 1, \cdots, p$ directly from the speech signal so as to obtain a good estimate of the spectral properties thereof.

Letting $\hat{a}_i$ denote the estimate of a_i we can define the *(forward) prediction error*

$$e(n) = s(n) - \sum_{i=1}^{N} \hat{a}_i s(n-i) \tag{4.1.4}$$

Similar to the forward prediction error we can define the *backward prediction error*

$$b(n) = s(n-N) - \sum_{i=1}^{N} \hat{a}_i s(n-i+1) \tag{4.1.5}$$

Now a good estimate can be found by a set of *predictor coefficients* that minimize the *(forward) mean-square prediction error*

$$E_p = E\left\{e^2(n)\right\} = E\left\{\left[s(n) - \sum_{i=1}^{N} \hat{a}_i s(n-i)\right]^2\right\} \tag{4.1.6}$$

over a short segment of speech. Similarly, with (4.1.28) we can define the backward mean-square prediction error. The resulting parameters are then assumed to be the parameters of the system function $H(z)$ as given in the model of speech production (4.1.2). The mean-square prediction error can be minimized by setting the partial derivatives of the mean-square prediction error with respect to the LP parameters equal to zero $\partial E_p / \partial \hat{a}_i = 0$ to arrive at p linear equations for the p unknown LP parameters $a_1, \ldots, a_p$:

$$\sum_{j=1}^{p} \hat{a}_j \phi_n(i,j) = \phi_n(i,0) \quad i = 1, \ldots, p \tag{4.1.7}$$

where

$$\phi_n(i,j) = E\left\{s(n-i)s(n-j)\right\} \tag{4.1.8}$$

In deriving (4.1.7), our major assumption was that the signal is stationary, which in the case of speech holds only for short segments of speech. As a consequence, we must replace the expectation of (4.1.8) by finite summations over a short length of speech samples:

$$\phi_n(i,j) = \sum_{m} s_n(m-i)s_n(m-j) \quad : i = 1, \ldots, p; \; j = 1, \ldots, p \tag{4.1.9}$$

There are two different ways of interpreting the last equation, leading to two different methods for estimating the prediction coefficients:

- the autocorrelation method
- the covariance method

These two methods along with lattice based methods, like the Burg method, are efficient and popular all-pole spectral estimation techniques. A detailed description of these methods follow in the next sections.

It is worth noting that in this section the least mean square approach was used to derive the equations for LP analysis. The maximum likelihood method can also be used. [Sri79].

4.1.2 The Autocorrelation Method (Levinson-Durbin Recursion)

The autocorrelation approach assumes that the segment $s_n(m)$ is zero outside the interval $0 \leq m \leq N-1$ where N is the length of the sample sequence. Introducing a finite length window $w(m)$ that is identically zero outside the given interval, the speech segment can be expressed as:

$$s_n(m) = s(m+n)w(m) \tag{4.1.10}$$

Assuming that the interest is only in the future prediction performance, the limits of (4.1.9) can be expressed as

$$\phi_n(i,j) = \sum_{m=0}^{N-1-(i-j)} s_n(m)s_n(m+i-j) \quad : i = 1, \ldots, p;\ j = 1, \ldots, p \tag{4.1.11}$$

The last can be rewritten as the *short-time autocorrelation function*:

$$\phi_n(i,j) = R(|i-j|) \quad : i = 1, \ldots, p;\ j = 1, \ldots, p \tag{4.1.12}$$

where

$$R(j) = \sum_{m=1}^{N-1-j} s_n(m)s_n(m+j) \tag{4.1.13}$$

Therefore (4.1.7) can be expressed as

$$\sum_{j=1}^{p} a_j R(|i-j|) = R(i) \quad : 1 \leq i \leq p \tag{4.1.14}$$

The set of equations (4.1.14), which are known as the *Yule-Walker equations*, can also be expressed as

$$\begin{bmatrix} R(0) & R(1) & \cdots & R(p-1) \\ R(1) & R(0) & \cdots & R(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ R(p-1) & R(p-2) & \cdots & R(0) \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ R(p) \end{bmatrix} \tag{4.1.15}$$

where the matrix is *Toeplitz*. That means the matrix is symmetrical and all the elements along a given diagonal are equal. Equation (4.1.15) can be solved by inverting the relevant $p \times p$ matrix, although this is usually not done due to its high computational cost and the associated accumulation of finite precision errors. In light of the special properties of a Toeplitz matrix, this inversion problem can be solved much more efficiently than is generally possible. The well known *Levinson-Durbin recursion*, which we now summarize, is designed to do exactly that. Set

$$a_0^{(0)} = 1; \qquad \epsilon_0 = R(0)$$

with the reflection coefficients

$$k_n = \frac{-1}{\epsilon_{n-1}} \sum_{i=1}^{n-1} R(i-n) a_i^{(n-1)} \qquad : n = 1 \leq i \leq N \tag{4.1.16}$$

where

$$a_i^{(n)} = \begin{cases} 1 & : i = 0 \\ a_i^{(n-1)} + k_n a_{n-i}^{*(n-1)} & : i = 1, \cdots, n-1 \\ k_n & : i = n \end{cases} \tag{4.1.17}$$

and the modelling error

$$\epsilon_n = \epsilon_{n-1}(1 - |k_{n|^2}) \tag{4.1.18}$$

After solving (4.1.16) to (4.1.18) recursively for $i = 1, \cdots, p$ the parameters are given by $a_i^{(N)}$ for $i = 1, \cdots, N$. An example can be found in [Kon94]

Key properties conducing to a preference for the autocorrelation method over other LP methods for speech processing applications are:

- **Computational Efficiency**

  Since the LP parameters are typically updated 100 times every second, algorithmic complexity is a key issue. As mentioned above, the simultaneous equations $\mathbf{R}_\mathrm{s} \cdot \mathrm{a} =$

$\mathbf{r}_s$ can be solved efficiently using the Levinson-Durbin. In addition, the reflection coefficients, to which we shall return in Section 4.3, are computed as a by-product of the Levinson-Durbin recursion.

- **Minimum-Phase Solution**
  The solution of the Yule-Walker equations guarantees that the prediction filter $H(z)$ is minimum-phase; i.e., all zeros fall within the unit circle. This implies that the LP synthesis filter $H(z)$ is stable.

4.1.3 The Covariance Method (Cholesky Decomposition)

The second approach commonly used to define the speech segment $s_n(m)$ and the limits on the summation is to fix the interval over which the mean-square error is computed

$$E = \sum_{m=0}^{N-1} e_n^2(m) \tag{4.1.19}$$

so that (4.1.7) can be rewritten as

$$\phi_n(i,j) = \sum_{m=-i}^{N-1-i} s_n(m)s_n(m+i-j) \quad : i = 1,\dots,p;\ j = 1,\dots,p \tag{4.1.20}$$

This equation is slightly different than (4.1.11), which was used to derive the autocorrelation method, inasmuch as it involves the interval $-i \leq m \leq N-1-i$ instead of $0 \leq m \leq N-1-(i-j)$. Actually (4.1.11) is not a true auto-correlation function, because it is approximated through a cross-correlation between two very similar finite length samples. Using (4.1.20), we can express (4.1.7) as

$$\sum_{j=1}^{p} a_j \phi_n(i,j) = \phi_n(i,0) \quad : 1 \leq i \leq p \tag{4.1.21}$$

or equivalently,

$$\begin{bmatrix} \phi_n(1,1) & \phi_n(1,2) & \cdots & \phi_n(1,p) \\ \phi_n(2,1) & \phi_n(2,2) & \cdots & \phi_n(2,p) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_n(p,1) & \phi_n(p-2) & \cdots & \phi_n(p,p) \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} \phi_n(1,0) \\ \phi_n(2,0) \\ \vdots \\ \phi_n(p,0) \end{bmatrix} \tag{4.1.22}$$

Although the differences between (4.1.14) and (4.1.21) appear to be minor, the solution of (4.1.21) is not as straightforward as the other case because ϕ is not Toeplitz. It is a symmetric positive definite matrix, however, and thus can be solved via the *Cholesky decomposition method*.

Let us express ϕ as

$$\phi = \mathbf{V}\mathbf{D}\mathbf{V}^T \tag{4.1.23}$$

where $\mathbf{V}$ is a lower triangular matrix whose main diagonal elements are all unity, and $\mathbf{D}$ is a diagonal matrix. The superscript T indicates matrix transposition. The elements of $\mathbf{V}$ and $\mathbf{D}$ are readily determined from (4.1.23) by solving for the $[i, j]^{\text{th}}$ element of both sides. An example can be found in [Rab78].

Because ϕ has the properties of a covariance matrix, this came to be known as the *covariance method*. This method does not guarantee the stability of the LP synthesis filter, nor is it computationally efficient for large p. Since the energy of the prediction error is minimized and the input speech signal is not windowed, however, the covariance method yields a residual signal with the highest achievable prediction gain.

The modified covariance method involves essentially the same steps as the covariance method. The final solution, however, is derived from the so-called partial correlations [Kay88], resulting in a minimum phase, and thus stable, LPC filter.

4.1.4 Basics of Lattice Methods

Both the autocorrelation and covariance methods involve two basic steps:

1. Computation of a matrix of correlation values $\phi_n(i, j)$

2. Solution of a set of linear equations

Although both steps can be performed efficiently, another class of auto-correlation methods has evolved, wherein the two steps are combined in a recursive manner. We now consider such *lattice* methods.

The basic idea of the lattice method (LM) for LP, which was first introduced by Itakura [Ita71], is that during calculation of the intermediate stages of the predictor parameters, the forward and the backward prediction error is considered. To see how the above mentioned steps are combined, we must recall the Levinson-Durbin recursion where the coefficients

$a_j^{(i)}, j = 1, 2, .., i$ are the optimal LPCs of an i^{th} order filter. Using these coefficients, the inverse filter can be written as

$$A^{(i)}(z) = 1 - \sum_{j=1}^{i} a_j^{(i)} z^{-j} \tag{4.1.24}$$

Using windowed segment of the signal $s_n(m) = s(n+m)w(m)$ as the input to this filter, the forward prediction error becomes

$$e^{(i)}(m) = s(m) - \sum_{j=1}^{i} a_j^{(i)} s(m-j) \tag{4.1.25}$$

and the backward prediction error becomes

$$b^{(i)}(m) = s(m-i) - \sum_{j=1}^{i} a_j^{(i)} s(m+j-i) \tag{4.1.26}$$

In the z-transform domain, this can be written as

$$E^{(i)}(z) = A^{(i)}(z)S(z) \tag{4.1.27}$$

and

$$B^{(i)}(z) = z^{-i}A^{(i)}(z^{-1})S(z) \tag{4.1.28}$$

By substituting (4.1.17) into (4.1.25), we obtain a recursion formula for $A^{(i)}(z)$ in terms of $A^{(i-1)}(z)$:

$$A^{(i)}(z) = A^{(i-1)}(z) - k_i z^{-i} A^{(i-1)}(z^{-1}) \tag{4.1.29}$$

Using (4.1.27) with (4.1.29), we obtain the z-transform of an i^{th} order prediction error from the $(i-1)^{\text{th}}$ order predictor error $E^{(i-1)}$ and a backward predictor error $B^{(i)}$

$$E^{i}(z) = \underbrace{A^{(i-1)}(z)S(z)}_{E^{(i-1)}(z)} - k_i \cdot \underbrace{z^{-i}A^{(i-1)}(z^{-1})S(z)}_{B^{(i)}(z)} \tag{4.1.30}$$

or

$$e^{(i)}(m) = e^{(i-1)}(m) - k_i \cdot b^{(i)}(m) \tag{4.1.31}$$

In light of (4.1.28) and (4.1.29), we may rewrite the backward prediction error as

$$B^{(i)}(z) = \underbrace{z^{-i}A^{(i-1)}(z^{-1})S(z)}_{B^{(i-1)}(z)} - k_i \cdot \underbrace{A^{(i-1)}(z)S(z)}_{E^{i-1}(z)} \tag{4.1.32}$$

or

$$b^{(i)}(m) = b^{(i-1)}(m-1) - k_i \cdot e^{(i-1)}(m) \tag{4.1.33}$$

Equations (4.1.31) and (4.1.33) clearly define the forward and backward prediction error of an i^{th} order predictor in terms of the corresponding prediction errors of order $(i-1)^{\text{th}}$. A prediction of order zero can be said to be no prediction at all, which implies

$$e^{(0)}(m) = b^{(0)}(m) = s(m) \tag{4.1.34}$$

This process can be depicted with the lattice network in Figure 4.1, which is a direct consequence of the Levinson-Durbin recursion (4.1.16) to (4.1.18), where the parameters k_i can be obtained. Relating these k_i parameters directly to the forward and backward prediction errors [Ita68] we find

$$k_i = \frac{\sum_{m=0}^{N-1} e^{i-1}(m) b^{(i-1)}(m-1)}{\sqrt{\sum_{m=0}^{N-1} (e^{(i-1)}(m))^2 \sum_{m=0}^{N-1} (b^{(i-1)}(m-1))^2}} \tag{4.1.35}$$

The last equation has the form of a normalized cross-correlation function; hence the k_i are known as *partial correlation coefficients* or *PARCOR coefficients*.

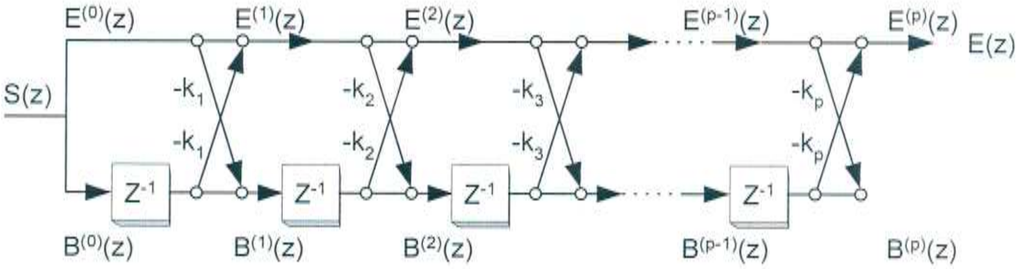


Figure 4.1: Block diagram of lattice methods

The calculation of the PARCOR coefficients (4.1.35) replaces (4.1.16) in the Levinson-Durbin recursion, and leads to an alternative method of matrix inversion. The coefficients thereby obtained are identical to those found with the autocorrelatation method. Both methods minimize the mean-squared forward prediction error.

The lattice approach is not limited to the Levinson-Durbin-type approaches and opens up for a whole new class of procedures based on the lattice configuration of Figure 4.1, e.g. Burg method (see next section), feed-back PARCOR ladder form and Makhoul covariance ladder algorithm. These and more approaches with self-contained algorithm summaries can be found in [Str90]. Also see [Hay02].

4.1.5 Burg Method

One lattice based approach is of particular interest, because it guaranties stability and minimizes the sum of the mean-squared *forward and backward prediction error*,

$$E_{\mathrm{fb}} = \sum_{m=0}^{N-1} \left[e^2(m) + b^2(m)\right] \tag{4.1.36}$$

This is known as the Burg Method [Rab78] or Burg's Harmonic Mean PARCOR Ladder Algorithm [Str90], after its discoverer.

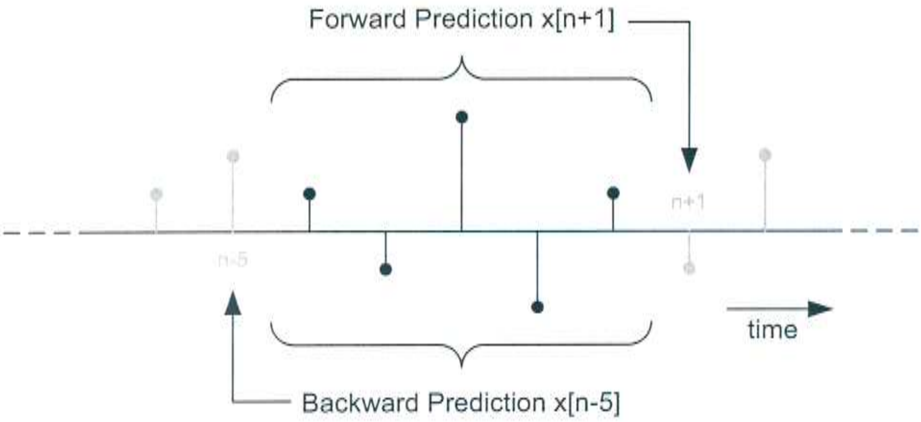


Figure 4.2: Illustration of forward and backward prediction at time n

There are actually two ways to derive the Burg method. One is through the harmonic mean of the forward and backward ladder reflection coefficients

$$\mathbf{k} = \frac{2\mathbf{k}_{\mathrm{e}} \cdot \mathbf{k}_{\mathrm{b}}}{\mathbf{k}_{\mathrm{e}} + \mathbf{k}_{\mathrm{b}}} \tag{4.1.37}$$

which will not further considered here and is only mentioned to give an explanation for the alternative name of this method. The second is through the aforementioned minimization of the forward and backward prediction error; see Figure 4.2. The desired minimum is obtained by differentiating E_{fb} as given in (4.1.36) with respect to k_i.

$$\begin{aligned} \frac{\partial E^i_{\text{fb}}}{\partial k_i} &= -2\sum_{m=0}^{N-1}\left[e^{(i-1)}(m) - k_i \cdot b^{(i-1)(m-1)}\right] b^{(i-1)}(m-1) \\ &\quad -2\sum_{m=0}^{N-1}\left[b^{(i-1)}(m-1) - k_i \cdot e^{(i-1)(m)}\right] e^{(i-1)}(m) \end{aligned} \tag{4.1.38}$$

Setting this derivative to zero and solving for k_i, we obtain *Burg's PARCOR coefficients*:

$$k_i = \frac{2\sum_{m=0}^{N-1} e^{i-1}(m) b^{(i-1)}(m-1)}{\sum_{m=0}^{N-1}(e^{(i-1)}(m))^2 + \sum_{m=0}^{N-1}(b^{(i-1)}(m-1))^2} \tag{4.1.39}$$

This can be expressed in vector form as

$$k_i = \frac{2(\mathbf{e}^{(i-1)})^T\mathbf{b}^{(i-1)}}{(\mathbf{e}^{(i-1)})^T\mathbf{e}^{(i-1)} + (\mathbf{b}^{(i-1)})^T\mathbf{b}^{(i-1)}} \tag{4.1.40}$$

A mayor advantage of the Burg method is that the coefficients k_i always satisfy $|k_i| \leq 1$, which guarantees a stable filter.

To prove this, we must recall the fundamental geometrical property $\mathbf{e}^T\mathbf{b} = |\mathbf{e}||\mathbf{b}|\cos\varphi \leq |\mathbf{e}||\mathbf{b}|$ and recognize that (4.1.40) is of the form

$$k_i = \frac{2|\mathbf{e}||\mathbf{b}|\cos\varphi}{|\mathbf{e}|^2 + |\mathbf{b}|^2} \tag{4.1.41}$$

It remains only to show that

$$\frac{2|\mathbf{e}||\mathbf{b}|}{|\mathbf{e}|^2 + |\mathbf{b}|^2} \leq 1 \tag{4.1.42}$$

It is sufficient to demonstrate that the difference between the denominator and the numerator above is greater than or equal to zero. This follows from the inequality

$$|\mathbf{e}|^2 + |\mathbf{b}|^2 - 2|\mathbf{e}||\mathbf{b}| = (|\mathbf{e}| - |\mathbf{b}|)^2 \geq 0 \tag{4.1.43}$$

4.1.6 Additional Linear Prediction Approaches

Although the previously discussed approaches are widely used and lead to good results, a variety of alternative solutions has appeared, some of which will be briefly discussed in this section.

- **Extended correlation matching:**
  The autocorrelation method only matches the first p correlations of the weighted speech signal with the impulse response of the synthesis filter while extended correlation matching is a weighted mean-square error match to $N_c \geq$ p correlations [Jac96]. A recursive procedure is necessary, and the minimum phase property does not hold in general.

- **Discrete all-pole modelling:**
  This is another iterative procedure that improves the spectral fit for segments corresponding to voiced speech. Introduced by El-Jaroudi and Makhoul [Jar91], this method fits an LP spectrum to a finite set of spectral points by minimizing a form of the Itakura-Saito distance measure. This is especially effective for the discrete line spectra exhibited in voiced speech. The improved spectral fit comes at the expense of possibly unstable synthesis filters.

- **Pole-zero methods:**
  Although pole-zero models can more accurately match the spectra of speech containing anti-resonances [Lim96], the computational complexity associated with these algorithms has been a compelling argument against their use in any real-time system. Obtaining the coefficients of pole-zero system typically involves a set of highly non-linear equations that must be solved iteratively. The Steiglitz-McBride algorithm [Jac96] is an example of such a method for finding a pole-zero fit.

  This methods is not of interest, as we shall confine our attention to spectral estimation techniques based on all-pole modeling.

4.1.7 High Resolution through Linear Prediction

The promising aspect of LP, and in particular the forward and backward LP, is that it makes more realistic assumptions concerning the values outside the measured interval which results in an extension of the measured autocorrelation sequence. This is apparent on comparing a) and b) with c) of Figure 4.3.

As a result, the window function can be eliminated without incurring any addtional distortions. This provides a higher resolution of the spectral estimate due to the fact that

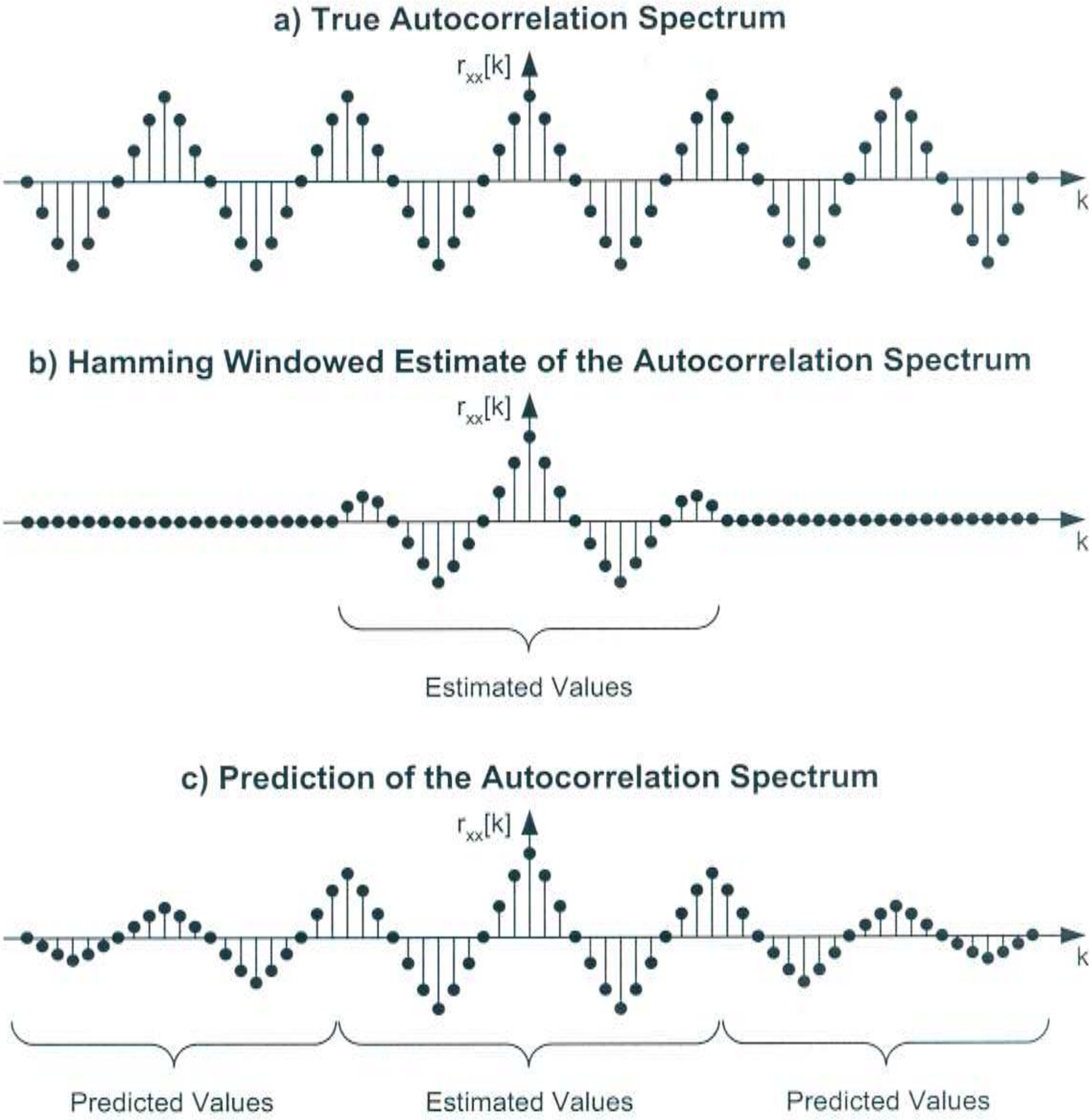


Figure 4.3: Illustrations of true and estimated autocorrelations

windowing reduces the available information [Kay88].

4.2 Interpretation of Linear Prediction in the Frequency-Domain

Up to this point, we have discussed LP methods in the time domain as represented by difference equations and correlation functions. Now we wish to relate these discussions to spectral estimation, our primary interest. The *power spectral density function* is defined as

$$S(\omega) = \sum_{n=-\infty}^{+\infty} \phi_n e^{-j\omega n}(n) \tag{4.2.1}$$

From this defintion, we see readily that the *power spectral density* (PSD) is the Fourier transform of the autocorrelation sequence

$$\phi_n = \mathbf{x}^T(t)\mathbf{x}(t-n) \tag{4.2.2}$$

Determining the PSD requires knowledge of the entire infinite autocorrelation sequence. But in practice, spectral estimation can only be performed on a finite observation interval. Hence, a contradiction exists between theory and practice. A convenient way out of this situation lies in the parameterization of the observed process.

Using the most general pole-zero model discussed in Section 4.1.1, the desired PSD is approximated by the squared absolute value of the pole-zero transfer function evaluated on the unit circle. Applying this approxmation, we obtain the *pole-zero-power spectral estimate*:

$$S_{\text{pole-zero}}(\omega) = \left| \frac{b_0 + b_1 e^{-j\omega} + b_2 e^{-2j\omega} + \cdots - b_p e^{-pj\omega}}{1 - a_1 e^{-j\omega} - a_2 e^{-2j\omega} - \cdots - a_q e^{-qj\omega}} \right|^2 \tag{4.2.3}$$

Any continuous PSD can be approximated to arbitrary precision by a pole-zero model whose numerator and denominator are of sufficiently high order.

This approach reduces the spectral estimation problem to a signal identification problem, where it is desired to determine the system parameters from a given set of data. The data set is then assumed to be the output of an unknown system with white Gaussian excitation modelled by a pole-zero filter.

Using an all-pole model, instead of the general model derived above, to get the *all-pole-power spectral estimation* of speech

$$S_{\text{all-pole}}(\omega) = \left| \frac{1}{1 - a_1 e^{-j\omega} - a_2 e^{-2j\omega} - \cdots - a_q e^{-qj\omega}} \right|^2 \tag{4.2.4}$$

has the advantage that it models the perceptually important spectral peaks better than the spectral valleys.

As derived, LP analysis can be viewed as a method of short-time spectrum estimation. It is widely used for this purpose in speech processing and coding, as well as having found numerous applications in other fields.

4.3 Alternative Representations of Spectral Parameters

So far we have shown that LP analysis produces an N-vector $\mathbf{a}$ of real prediction coefficients, which represents an optimal spectral estimate of a windowed speech signal. It is also possible to represent this set of N real predictor coefficients in other formats, some of which are more useful for applications like speech recognition, or can be more easily interpreted than others. In this chapter, two commonly used representations are given. More representations can be found in [Rab78]

4.3.1 Reflection and PARCOR Coefficients

Unlike the prediction coefficients, the reflection coefficients k_i guarantee stability provided $|k_i| \leq 1$ for all $1 \leq i \leq N$. This fact is useful when implementing LP filters whose values are interpolated over time.

The reflection coefficients can be calculated from the LP using a backward recursion of the form

$$k_i = a_i^{(i)} \qquad : i = p, \ldots, 1 \tag{4.3.1}$$

$$a_j^{(i-1)} = \frac{a_j^{(i)} + a_i^{(i)} \cdot a_{i-j}^{(i)}}{1 - k_i^2} \qquad : 1 \leq j < i \tag{4.3.2}$$

where we initialize $a_i^p = a_i$.

The negatives of the reflection coefficients are called partial correlation, or PARCOR, coefficients.

4.3.2 LP Cepstrum

The cepstrum[1] [Rab78] is the inverse Fourier transform of the logarithm of the magnitude spectrum of a signal:

$$c_n = \frac{1}{2\pi} \int\limits_{-\pi}^{\pi} \log_e H(e^{j\omega}) e^{j\omega n} d\omega \tag{4.3.3}$$

where

$$H(e^{j\omega}) = \frac{G}{1 - \sum_{k=1}^{p} a_k e^{-j\omega k}} \tag{4.3.4}$$

and the complex logarithm $\log_e(e^{j\omega}) = \ln|e^{j\omega}| + j\theta(\omega)$ is used. This can be shown [Hua01] to simplify to the following recursion

$$h[n] = \begin{cases} 0 & : n < 0 \\ \ln G & : n = 0 \\ a_n + \sum_{k=1}^{n-1} \frac{k}{n} h[k] a_{n-k} & : 0 < n \leq p \\ \sum_{k=n-p}^{n-1} \frac{k}{n} h[k] a_{n-k} & : n > p \end{cases} \tag{4.3.5}$$

Note that although the number of LP cepstral coefficients (LPCC) are limited, the number of cepstral coefficients are infinite. Using the LPCC as direct input in a speech recognition system a number between 12 and 20 has to be empirically shown to be significant [Hua01].

To adapt several concepts of the psychology of hearing to the LPCC approach Hermansky [Her90] introduced the *perceptual linear predictive* (PLP).

More details about cepstral coefficients in general can be found in [Opp89].

4.4 LP-Based MFCC Speech Recognition Frontend

The LP-based MFCC approach is basically the same as the FFT-based MFCC approach. But the MFCC are derived from a SE based on a reconstructed PS of the LPC instead of a PS based on the Fourier transform, also compare Figure (3.3) with Figure (4.4).

[1] In 1963 the word cepstrum was first used by Bogert, Healy and Tukey which is derived from the interchange of the first two characters in the word spectrum. They called it cepstrum because "in general, we find ourself operation on the frequency side in ways customary on the time side and vice versa".

A comparison of LP- and FFT-based MFCC for noise robust ASR can be found in [Wet01]. There is shown that the LP approach outperforms their FFT counterpart in most of the investigated adverse acoustic conditions, intraocular in *mismatched acoustic conditions*.

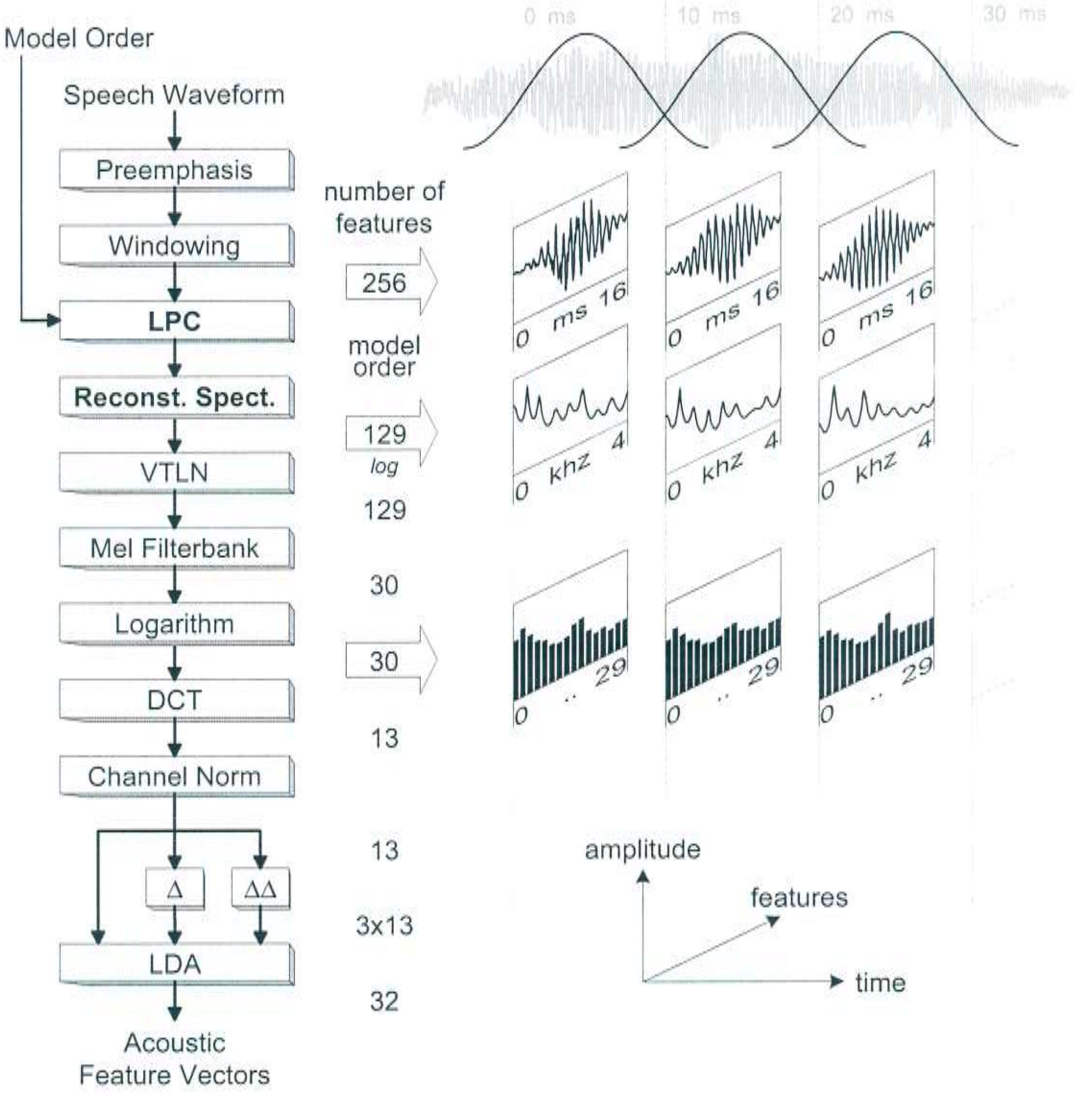


Figure 4.4: LP-based MFCC speech recognition frontend

5 Minimum Variance Spectral Estimation

All-pole filters based on LP methods lead to filters that model unvoiced speech well and perform relatively well for low-pitched voiced speech. But in the case of medium- and high-pitched voiced speech, they do not provide a good spectral estimate [Jar91]. This is because LP based all-pole filters tend to envelope the spectrum as tightly as possible, and will under certain conditions descend down to the level of residual noise in the gap between two *harmonic partials* (sounds with a prevalent partial structure). This will happen whenever the space between partials is large, as in high pitched sounds, and when the order is high enough;i.e., there are enough poles to cover every partial peak [Sch98]. Therefore, filters which yield SE with a smoother contour than LP based all-pole filters are desired. Murthi and Rao [Mur97] proposed the *minimum variance distortionless response* (MVDR) as a spectral estimation technique for speech, and demonstrated that it provides superior modelling for medium- and high-pitched voices.

5.1 Theoretical Background

The *minimum variance spectral estimation* (MVSE) was originally introduced by Capon [Cap69] and is also known as *Capon's method* [Li98] or the *maximum-likelihood method*[1] [Mus85]. It belongs to the class of filterbank approaches, which are in general implemented in the following steps:

1. Passing of the observed signal through a bandpass filter with a variable *frequency of interest* (FOI) ω_{foi}.

[1] As this formula has no bearing on the classical principle of maximum likelihood and the expression MVDR, which owes its origin to Owsley [Hay91], is commonly used in the recent literature [Hay91][Mur00][Dha01], we prefer to use the terminology MVDR, which in the case of spectral estimation is shortened to MVSE which is also used in Kays book on spectral estimation [Kay88].

2. Measuring of the output power of the filter.
3. Calculating of the spectral power estimate by dividing the measured power by the bandwidth of the filter.

Hence, spectral estimation, from the viewpoint of filterbank analysis, is a problem of filter design subject to some specific constraints [Lag84]. For MVDR, the relevant constraint is the *distortionless constraint*, which can be stated as: *The signal at the frequency of interest* ω_{foi} *must pass undistorted (i.e., unity gain).*

$$H(e^{j\omega_{\text{foi}}}) = \sum_{k=0}^{M} h^*(k)e^{-j\omega_{\text{foi}}k} = 1 \tag{5.1.1}$$

In vector form this simplifies to

$$\mathbf{s}^H(\omega_{\text{foi}}) \cdot \mathbf{h^*}_{\text{foi}} = 1 \tag{5.1.2}$$

where $\mathbf{s}(\omega)$ is the *fixed frequency vector*

$$\mathbf{s}(\omega) = (1, e^{-j\omega}, \ldots, e^{-jM\omega})^T \tag{5.1.3}$$

and $\mathbf{h}_{\text{foi}} = (h(0), h(1), \ldots, h(M))^T$.

The distortionless filter $\mathbf{h}_{\text{foi}}$ can now be obtained by the *constrained minimization problem* which minimizes the output power of the overall frequency domain:

$$\min_{\mathbf{h}_{\text{foi}}} \mathbf{h}_{\text{foi}}^H \phi_{M+1} \mathbf{h}_{\text{foi}} \quad \text{subject to} \quad \mathbf{s}^H(\omega_{\text{foi}})\mathbf{h}_{\text{foi}} = 1 \tag{5.1.4}$$

In the above, ϕ_{M+1} is the $(M+1) \cdot (M+1)$ Toeplitz autocorrelation matrix of the filter output,

$$y(i) = \sum_{l=0}^{M} h^*(l)u(i-l) \tag{5.1.5}$$

where $h_0, h_1, \ldots, h_M$ are the transversal filter coefficients and $\mathbf{u}$ denotes the inputs of the filter.

In order to solve this constrained minimization problem, let us define the *constrained cost function*:

$$E = \underbrace{\sum_{i=M+1}^{N} |y(i)|^2}_{\text{output energy}} + \underbrace{\lambda\left(\sum_{k=0}^{M} h^*(k)e^{-j\omega_{\text{foi}}k} - 1\right)}_{\text{linear constraints}} \tag{5.1.6}$$

where λ denotes the complex *Lagrange multiplier* [Hay91]. At the optimum, the gradient vector ∇E must be zero. Thus, we find the k^{th} element of the gradient vector of (5.1.6) as

$$\nabla_k E = 2 \sum_{i=M+1}^{N} u(i-k)y^*(i) + \lambda^* e^{-j\omega_{\text{foi}}k} \tag{5.1.7}$$

Substituting (5.1.5) into (5.1.7) and rearranging the terms, we may write

$$\nabla_k E = 2 \sum_{l=0}^{M} h(l) \sum_{i=M+1}^{N} u(i-k)u^*(i-l) + \lambda^* e^{-j\omega_{\text{foi}}k} \tag{5.1.8}$$

Given the time average on $\sum_{l=0}^{M} h(l) \sum_{i=M+1}^{N} u(i-k)u^*(i-l)$, we may represent the *time averaged autocorrelation function* of tap inputs as $\phi(l,k) = \sum_{i=M+1}^{N} u(i-k)u^*(i-l)$ and thus, we can replace (5.1.8) by

$$\nabla_k E = 2 \sum_{l=0}^{M} h^*(l)\phi(l,k) + \lambda^* e^{-j\omega_{\text{foi}}k} \tag{5.1.9}$$

With $\nabla_k E = 0$ for $k = 0, 1, \ldots, M$ to minimize the constrained cost function E we find from (5.1.9) that the *tap-weights* of the optimized transversal filter has to satisfy the $M+1$ simultaneous equations

$$\sum_{l=0}^{M} \hat{h}(l)\phi(l,k) = -\frac{1}{2}\lambda^* e^{-j\omega_{\text{foi}}k} \qquad : k = 0, 1, \ldots, M \tag{5.1.10}$$

Rewriting this in matrix notation we get:

$$\phi\hat{\mathbf{h}} = -\frac{1}{2}\lambda^* s(\omega_{\text{foi}}) \tag{5.1.11}$$

Under the assumption that the time-averaging correlation matrix ϕ is nonsingular and therefore its inverse ϕ^{-1} exists we may solve (5.1.11) for the optimum tap-weight vector

$$\hat{\mathbf{h}} = -\frac{1}{2}\lambda^* \phi^{-1}\mathbf{s}(\omega_{\text{foi}}) \tag{5.1.12}$$

Now, only the evaluation of the Lagrange multiplier λ remains and can be solved by the use of the distortionless constrained of (5.1.2), evaluations of the inner product of the vectors s_{foi} and $\hat{\mathrm{h}}$ in (5.1.12) and setting the results equal to unity and solving for λ.

$$\lambda^* = -\frac{2}{\mathbf{s}^H(\omega_{\text{foi}})\phi^{-1}\mathbf{s}(\omega_{\text{foi}})} \tag{5.1.13}$$

Finally, we obtain the *MVDR estimate* of the *tab-weight vector* by substituting this value of λ into (5.1.12)

$$\hat{\mathbf{h}} = \frac{\phi^{-1}\mathbf{s}(\omega_{\text{foi}})}{\mathbf{s}^H(\omega_{\text{foi}})\phi^{-1}\mathbf{s}(\omega_{\text{foi}})} \tag{5.1.14}$$

For further simplification, let $S_{\text{MV}}(\omega_{\text{foi}})$ denote the minimum value of the output energy E which can be expressed as

$$S_{\text{MV}}(\omega_{\text{foi}}) = \hat{\mathbf{h}}^H\phi\hat{\mathbf{h}} \tag{5.1.15}$$

and with its substitution into (5.1.14) we get

$$S_{\text{MV}}(\omega_{\text{foi}}) = \frac{1}{\mathbf{s}^H(\omega_{\text{foi}})\phi^{-1}\mathbf{s}(\omega_{\text{foi}})} \tag{5.1.16}$$

Recalling the definition of the *fixed frequency vector* [2]

$$\mathbf{s}(\omega) = [1, -e^{j\omega}, \dots, e^{-jM\omega}]^T \tag{5.1.17}$$

and replacing the frequency of interest ω_{foi} by the variable ω discovering the whole frequency-band of interest we get a more general interpretation called the *Minimum Variance Spectrum Estimation*:

$$S_{\text{MV}}(\omega) = \frac{1}{\mathbf{s}^H(\omega)\phi^{-1}\mathbf{s}(\omega)} \tag{5.1.18}$$

5.2 Fast MVDR Spectrum Computation

In the last section we have derived the principles of the MVDR spectrum. In this section we derive a fast algorithm for computing the MVDR spectrum, due to Musicus [Mus85], which exploits the Toeplitz property of the correlation matrix.

Recall (5.1.18) where the MVSE is defined in terms of the inverse matrix ϕ^{-1} of the $(N+1)\cdot(N+1)$ Hermitian Toeplitz correlation matrix ϕ with the $[l,k]^{\text{th}}$ element $R_{l,k} = R(k-l)$ under the assumption that the matrix is positive definite and thus invertible. If $R_{l,k}^{-1}$ denotes the $[l,k]^{\text{th}}$ element of ϕ^{-1}, we may rewrite (5.1.18) in the form

$$S_{\text{MV}}(\omega) = \frac{1}{\sum_{k=-M}^{M}\mu(k)e^{-j\omega k}} \tag{5.2.1}$$

[2] Also known as the frequency-scanning vector [Hay91]

where

$$\mu(k) = \sum_{l=\max(0,-k)}^{\min(N-k,N)} \phi^{-1}_{l,l+k} \tag{5.2.2}$$

Let the order N prediction error $\epsilon = \epsilon_N$ and prediction coefficients

$$\mathbf{a}^{(N)} = [a_0^{(N)}\ a_1^{(N)}\ \cdots\ a_N^{(N)}]^T$$

where $a_0^{(N)} = 1$ satisfy

$$\phi \begin{bmatrix} 1 \\ a_1^{(N)} \\ \vdots \\ a_N^{(N)} \end{bmatrix} = \begin{bmatrix} \epsilon \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{5.2.3}$$

Then the *Gohlberg-Semencul formula* [Goh72, Kai78] states that the $[l,k]^{\text{th}}$ entry of a Toeplitz matrix ϕ^{-1} can be written as

$$\phi^{-1}_{l,k} = \frac{1}{\epsilon}\sum_{i=0}^{l} a_i^{(N)} a_{i+(k-l)}^{*(N)} - a_{N+1-i}^{*(N)} a_{N+1-i-(k-l)}^{(N)} \qquad : k \geq l \tag{5.2.4}$$

Substituting (5.2.4) into (5.2.2) with restriction to $k \geq 0$, we get

$$\frac{1}{\epsilon}\sum_{l=0}^{N-k}\sum_{i=0}^{l} a_i^{(N)} a_{i+k}^{*(N)} - \frac{1}{\epsilon}\sum_{l=0}^{N-k}\sum_{i=0}^{N-k} a_{N+1-i}^{*(N)} a_{N+1-i-k}^{(N)} \tag{5.2.5}$$

Interchanging the order of summation and setting $j = N+1-i-k$ we can rewrite $\mu(k)$ as

$$\mu(k) = \frac{1}{\epsilon}\sum_{i=0}^{N-k}\sum_{l=i}^{N-k} a_i^{(N)} a_{i+k}^{*(N)} - \frac{1}{\epsilon}\sum_{j=1}^{N+1-k}\ \sum_{l=N+1-j-k}^{N-k} a_{j+k}^{(N)} a_{i+k}^{(N)} \tag{5.2.6}$$

The terms that do not involve the index l permit us to collapse the summation over l into a multiplicative integer constant. Thus, the summations can then be combined. Moreover, using the Levinson-Durbin recursion, or any other method for solving the LP problem of the prediction-error filter of (5.2.3), we may now formulate a fast algorithm for computing the MVDR spectrum as follows:

1. Calculate the LPC as given in (4.1.16) to (4.1.18).

2. Compute the parameters

$$\mu_k = \begin{cases} \sum_{i=0}^{N-k}(N+1-k-2i)a_i^{(N)} a_{i+k}^{*(N)} & : k = 0,\cdots,N \\ \mu^*_{-k} & : k = -N,\cdots,-1 \end{cases} \tag{5.2.7}$$

3. Compute the MVDR spectrum according to

$$S_{\mathrm{MV}}(\omega) = \frac{\epsilon}{\sum_{k=-M}^{M} \mu_k e^{-j\omega k}} \tag{5.2.8}$$

5.3 Scaled Minimum Variance Spectral Estimation

Spectral peaks have been shown to be particular robust to additive noise in the logarithmic domain, since $\log(a+b) \approx \log(\max\{a,b\})$ [Bar97]. Therefore we propose to match the MVDR derived spectrum to the highest spectral peak.

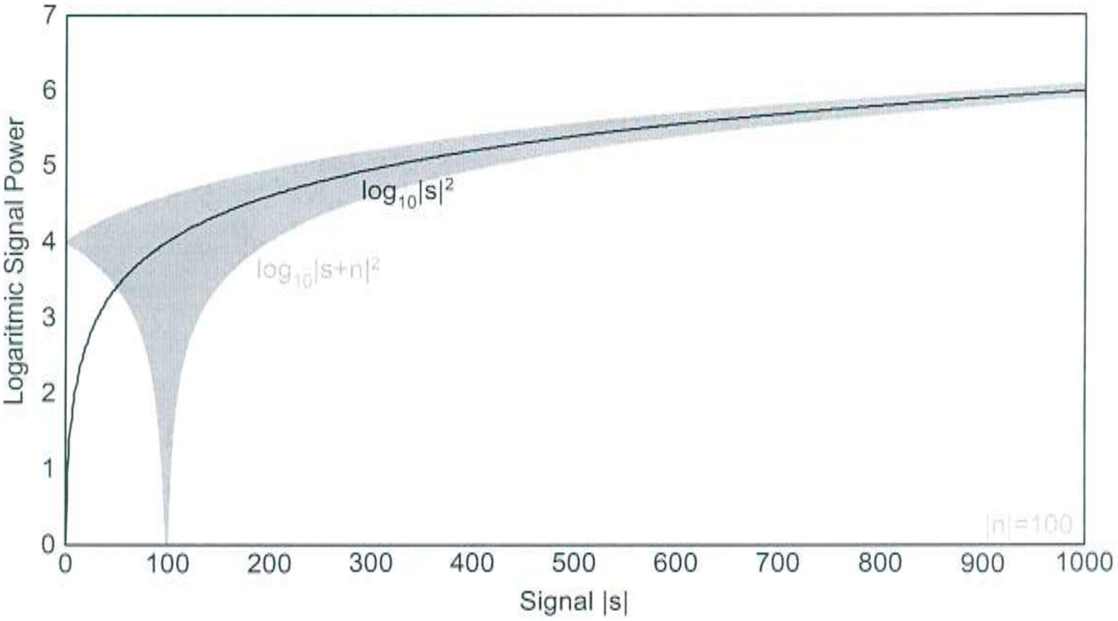


Figure 5.1: Logarithmic power of a signal disturbed by additive noise.

To fully understand why the suggested scaling should be useful, we must first investigate how additive noise influences the features in the LPS. Define the logarithmic domain signals

$$S_{\log} = \log(|s|^2) \tag{5.3.1}$$
$$\hat{S}_{\log} = \log(|\hat{s}|^2) \tag{5.3.2}$$

For additive noise n we may write

$$S_{\log} + D_{\log} = \log(|s+n|^2) \tag{5.3.3}$$

$$\hat{S}_{\log} + \hat{D}_{\log} = \log(|\hat{s} + n|^2) \tag{5.3.4}$$

where $D_{\log}$ and $\hat{D}_{\log}$ denote the logarithmic power differences between the noise free and noisy signals. Now we can solve the equations for $D_{\log}$ and $\hat{D}_{\log}$

$$D_{\log} = 2\log|s+n| - 2\log|s| = 2\log\left|\frac{s+n}{s}\right| = 2\log\left|1+\frac{n}{s}\right| \tag{5.3.5}$$

$$\hat{D}_{\log} = 2\log|\hat{s}+n| - 2\log|\hat{s}| = 2\log\left|\frac{\hat{s}+n}{\hat{s}}\right| = 2\log\left|1+\frac{n}{\hat{s}}\right| \tag{5.3.6}$$

Assuming $|n| < |s|$, we wish to prove that

$$|\hat{s}| > |s| \Rightarrow |\hat{D}_{\log}| < |D_{\log}| \tag{5.3.7}$$

This can be established with the following chain of inequalities:

$$\begin{array}{lcl}
 & |\hat{s}| > |s| & \\
\Rightarrow & 1 + \frac{|n|}{|s|} > 1 + \frac{|n|}{|\hat{s}|} & \\
\Rightarrow & \left|1+\frac{n}{s}\right| > \left|1+\frac{n}{\hat{s}}\right| & : |n| < |s| \\
\Rightarrow & \begin{array}{r} 2\log\left|1+\frac{n}{s}\right| > 2\log\left|1+\frac{n}{\hat{s}}\right| \\ -2\log\left|1+\frac{n}{s}\right| > -2\log\left|1+\frac{n}{\hat{s}}\right| \end{array} & \begin{array}{l} : |n| < |s| \wedge \frac{n}{s} > 0 \\ : |n| < |s| \wedge \frac{n}{s} < 0 \end{array} \\
\Rightarrow & \left|2\log\left|1+\frac{n}{s}\right|\right| > \left|2\log\left|1+\frac{n}{\hat{s}}\right|\right| & : |n| < |s| \\
\Rightarrow & \left|D_{\log}\right| > \left|\hat{D}_{\log}\right| & : |n| < |s|
\end{array}$$

This result is also apparent from Figure 5.1, where the grey plane is getting smaller to the right for $|s| > |n|$. Given (5.3.7), we can calculate the effects on the LSD as follows:

$$\left|\hat{D}_{\log}\right| < \left|D_{\log}\right| \tag{5.3.8}$$

This can be extended as

$$\left|\hat{S}_{\log} - (\hat{S}_{\log} + \hat{D}_{\log})\right| < \left|S_{\log} - (S_{\log} + D_{\log})\right| \tag{5.3.9}$$

and subsequently rewritten as

$$\left|\log\left(|\hat{s}|^2\right) - \log\left(|\hat{s}+n|^2\right)\right| < \left|\log\left(|s|^2\right) - \log\left(|s+n|^2\right)\right| \tag{5.3.10}$$

	SNR	max(FFT)	max(SFFT)	max(MVDR) autocorr.	max(MVDR) Burg
variance	*10 dB*	+ 0.0065	+ 0.0062	+ 0.0154	+ 0.0158
	8 dB	+ 0.0083	+ 0.0081	+ 0.0189	+ 0.0193
	6 dB	+ 0.0103	+ 0.0098	+ 0.0225	+ 0.0230
	4 dB	+ 0.0118	+ 0.0117	+ 0.0253	+ 0.0268
	2 dB	+ 0.0148	+ 0.0145	+ 0.0301	+ 0.0306
	*robot noise**	+ 0.0206	+ 0.0196	+ 0.0264	+ 0.0268
bias	*10 dB*	+ 0.0359	+ 0.0103	+ 0.0549	+ 0.0415
	8 dB	+ 0.0306	+ 0.0176	+ 0.0562	+ 0.0429
	6 dB	+ 0.0508	+ 0.0256	+ 0.0563	+ 0.0431
	4 dB	+ 0.0570	+ 0.0320	+ 0.0556	+ 0.0412
	2 dB	+ 0.0685	+ 0.0439	+ 0.0521	+ 0.0392
	*robot noise**	+ 0.0962	+ 0.0702	- 0.0770	- 0.0882

Table 5.1: Distortions of maxima of logarithmic power spectrum (* SNR 5dB).

Now lets denote $\hat{s} \equiv \max s(m)$ and $\hat{n} \equiv n(\text{argmax}[s(m)])$, which means that the noise is taken at the frequency where the highest spectral peak occurs. We may generalize‘(5.3.10) as

$$\left| \log\left(|\hat{s}|^2\right) - \log\left(|\hat{s}+\hat{n}|^2\right) \right| < \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left[\log\left(|s(m)|^2\right) - \log\left(|s(m)+\hat{n}|^2\right)\right]^2} \quad (5.3.11)$$

assuming that the noise is equally distributed where

$$S(m) = |s(m)|^2 \quad (5.3.12)$$

$$S_{\text{noise}}(m) = |s(m) + E\{n\}|^2 \quad (5.3.13)$$

$$\hat{S} = |\hat{s}|^2 \quad (5.3.14)$$

$$\hat{S}_{\text{noise}} = |\hat{s} + E\{n\}|^2 \quad (5.3.15)$$

denote the expected power values of the noise free and noisy signals. We may then further generalize as

$$\left|\log\left(\hat{S}\right) - \log\left(\hat{S}_{\text{noise}}\right)\right| < \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left[\log\left(S(m)\right) - \log\left(S_{\text{noise}}(m)\right)\right]^2} \tag{5.3.16}$$

The left hand side of (5.3.16) is the expected LSD at the highest amplitude in noise while the right hand side is the expected LSD over a given frequency band. Thus, we just have proved that the LSD is smaller at the highest amplitude than the expected LSD averaged over all considered frequencies.

This theoretical result can be verified by comparing the highest point of the FFT and smoothed FFT (SFFT) spectra to the highest point of the MVDR spectral envelope which uses the prediction error variance ϵ for scaling. Comparing the figures of Table 5.1, we clearly see that the variance of the MVDR approach, caused by additive noise, is much higher.

Therefore, we can increase our expected LSD, caused by additive noise, by *scaling* the amplitude of the envelope to match the highest point in the spectrum to the highest point of the spectrum defined by the Fourier transform. And not, as is commonly done, by considering *all* frequency bands to calculate the height of the SE. In the case of MVDR, this novel method should has been dubbed *scaled minimum variance distortionless response* (SMVDR). A schematic illustrating the derivation of the scaled SE using the MVDR approach is given in Figure 5.2.

Further insight into this phenomenon can be obtained by visualizing how additive noise influences the log-spectral features (LSF)s. Thus, let us plot the undisturbed energies of the LPS on the x-axis and the disturbed energies of the LPS on the y-axis. The black line in Figure 5.3 a) shows the ideal case of a noise free speech signal; here all points fall on the line $x = y$. In the case of additive noise (see gray line), the lower values of the PS are lifted to higher energies; i.e., the low-energy components are *masked* by noise and their information is lost, which results in *missing features*. To cope with this problem, *missing feature theory* [Coo97, Lip97] was developed.

Comparing Figure 5.3 b) with c), the problem which follows upon using the conventional

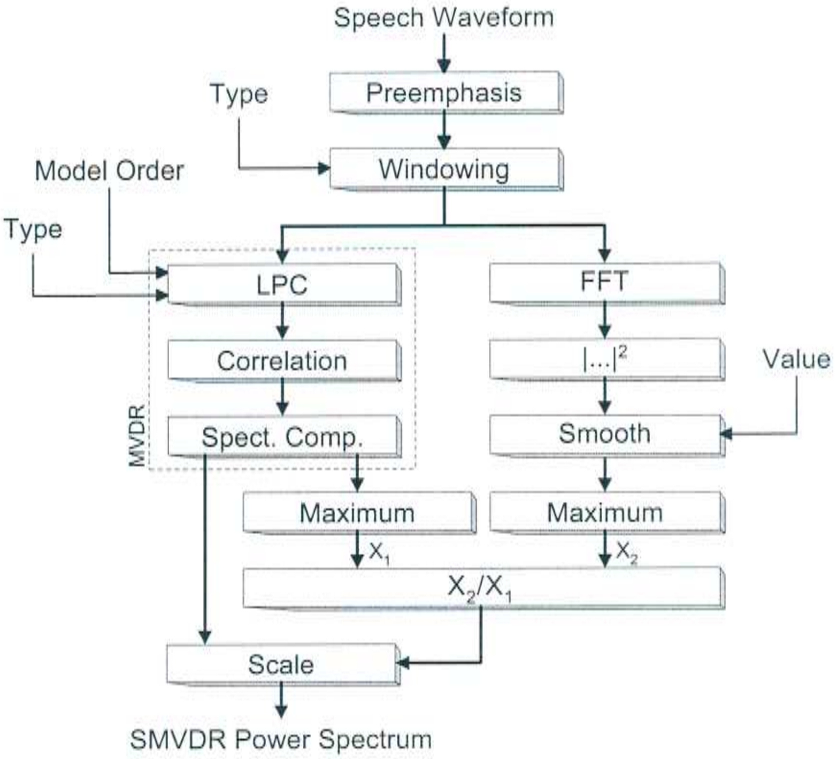


Figure 5.2: Block diagram of the SMVDR derived envelope

MVDR in the presence of additive noise is readily apparent: Because of the high variance of the noisy signal in the MVDR approach, there is a broad band instead of narrow ribbon, even in the high energy regions. As proposed, the scaling of the maximum of the MVDR spectrum to the spectral peak of the FFT to obtain the SFFT spectrum results in a decrease of spectral distortion. Thereby the SFFT provides more useful features and has fewer missing features than both conventional MVDR, which is clear upon comparing Figure 5.3 b) with d), and FFT spectral estimates, which can be seen by comparing Figure 5.3 c) with d).

Comparing the LSF in Figure 5.3 c) and d) with the Mel-filtered LSF of Figure 5.4, we see that the reduction of the features provided by the Mel filterbank from commonly 256 to 30 results in a reduction of missing features in the FFT case and a further reduction in the SMVDR case.

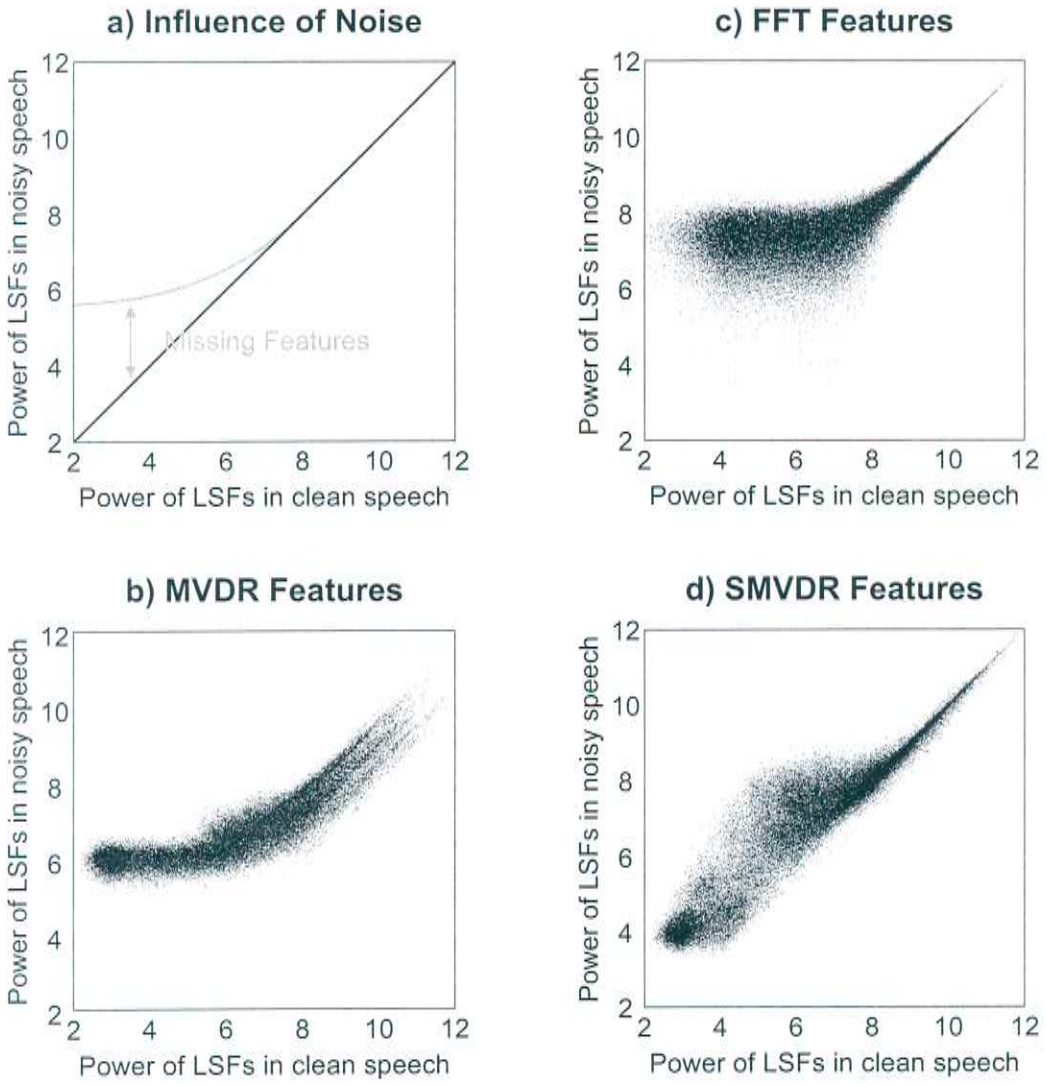


Figure 5.3: a) LSFs of a noise free signal (black line) and a disturbed feature (gray line). b-d) LSF disturbance of FFT, MVDR and SMVDR at SNR = 8 dB.

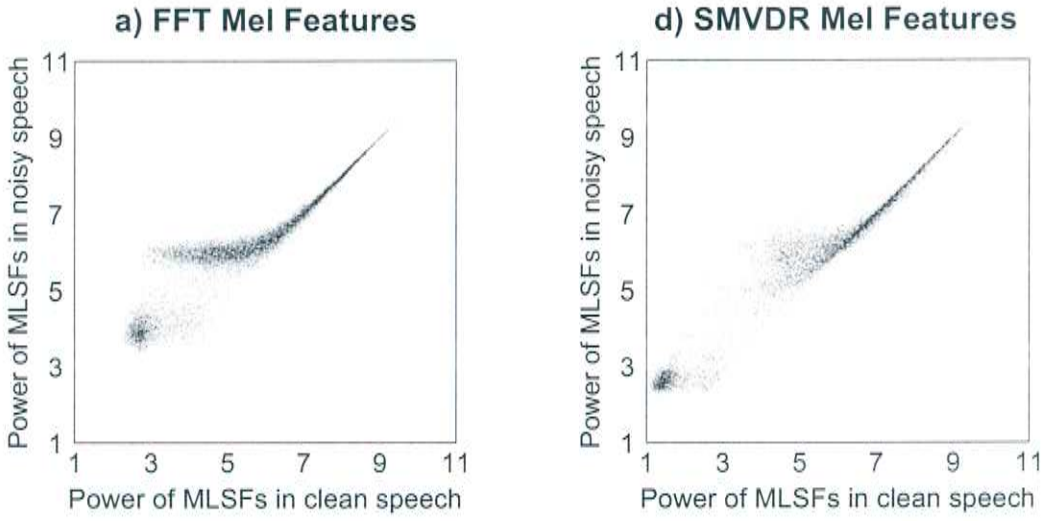


Figure 5.4: *Mel-filtered LSF* (MLSF) of FFT and SMVDR Spectrum at SNR = 8 dB.

To compare the LSD of one speech utterance using different spectral estimates, we must normalize the LSD by dividing the logarithmic clean speech energy

$$\text{normalized LSD} = \frac{\sqrt{\frac{1}{M}\sum_{m=1}^{M}[\log_{10} S(m) - \log_{10} \hat{S}(m)]^2}}{\sum_{m=1}^{M} \log_{10} S(m)} \tag{5.3.17}$$

because different approaches of spectral estimates results in different energies.

Comparing b) with c) and d) with e) of Figure 5.6, we see that the scaling of the MVDR spectra reduces the variance significantly. The same conclusion follows upon examining Table 5.2.

As mentioned in Section 3.4, in the case of voiced and transitional sounds, the spectral variation within the same utterance is higher using the Fourier spectrum in comparison to the SE. This should be demonstrated here in more detail. Therefore, plots of SEs are given in Figure 5.5 as a comparison to the Fourier spectrum of Figure 3.6. As well as Table 5.3 of the normalized LSD over consecutive frames which states the claimed.

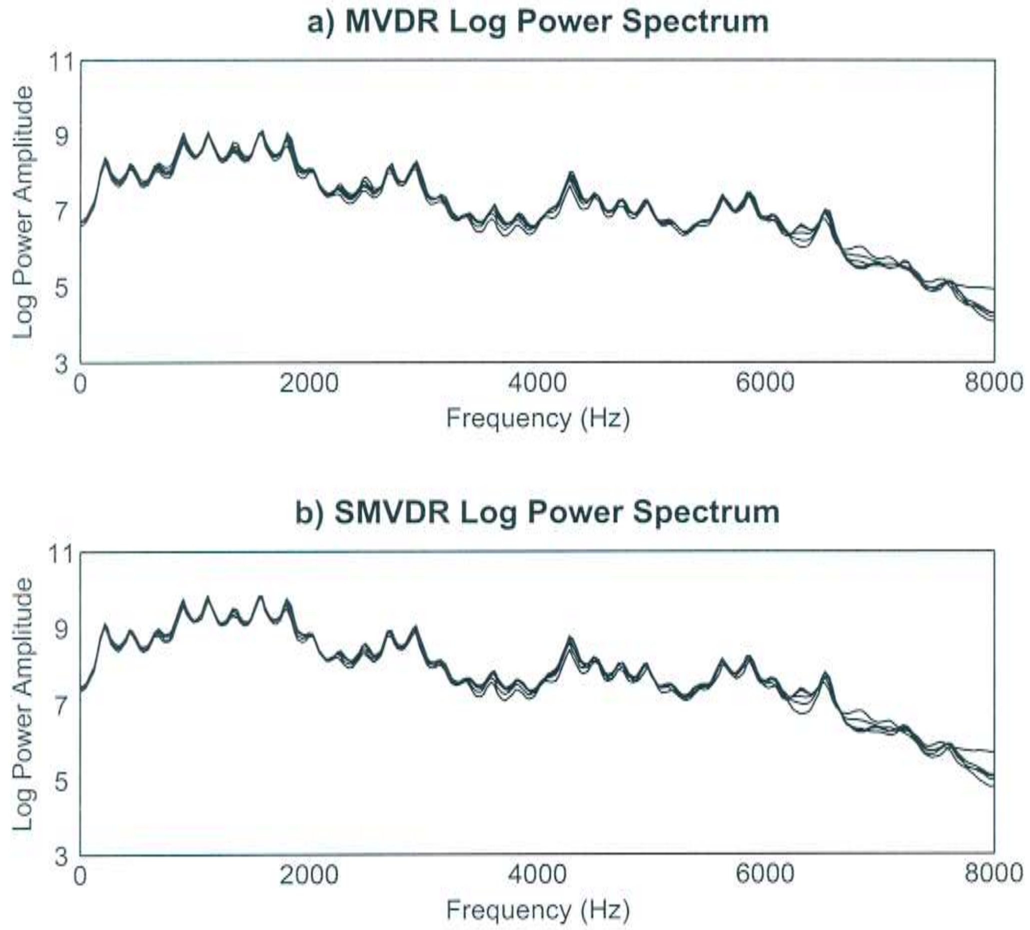


Figure 5.5: Power spectrum of 5 consecutive frames (2ms steps) over a stationary part

LPC	*spectrum*	*bias*	*variance*
	FFT	1.3318	3.6766
autocorr.	*MVDR*	1.447	7.6121
	SMVDR	1.3262	3.5937
Burg	*MVDR*	1.0541	7.1795
	SMVDR	1.3529	4.3663

Table 5.2: Comparison of bias and variance of the normalized LSF over all frames in noise SNR = 2 dB using different spectra. ((S)MVDR of order 120)

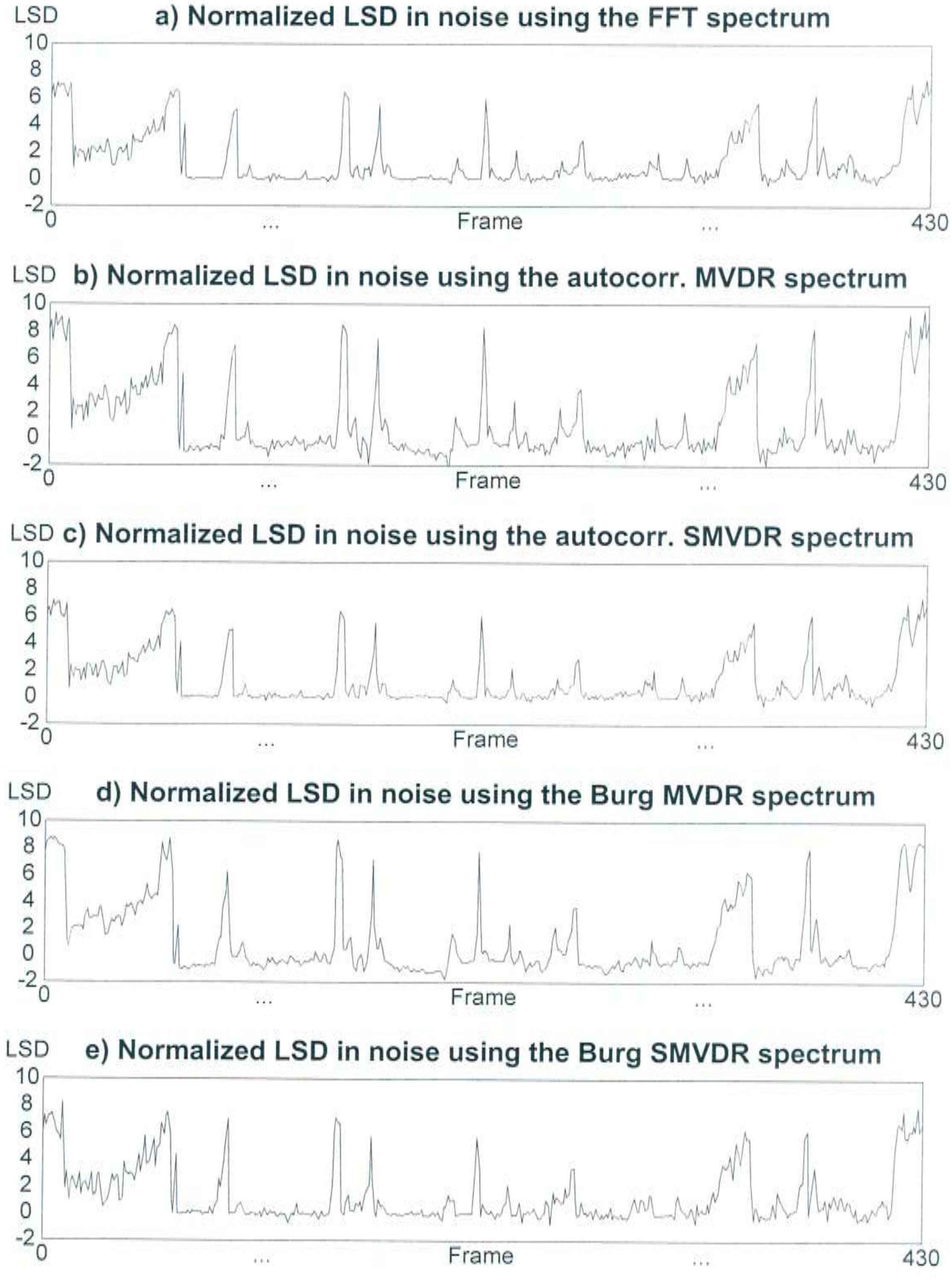


Figure 5.6: Adjusted LSD over frames in noise SNR = 2 dB using different spectra. ((S)MVDR of order 120)

Frame	FFT	MVDR	SMVDR
1-2	2.444	0.831	0.735
2-3	2.670	0.368	0.379
3-4	1.791	0.306	0.280
4-5	2.029	0.412	0.513
average	2.234	0.479	0.477

Table 5.3: Comparison of LSD (normalized energy, all values are multiplied by 1000) of 5 consecutive frames (2ms steps) over a stationary part.

5.4 Warped Minimum Variance Spectral Estimation

In the conventional form, the MVDR of a sample is obtained from the correlation of previous elements, in the case of the autocorrelation method, where the z-transform of the synthesis filter is given as

$$S_{\mathrm{MV}}(z) = \frac{\epsilon}{\sum_{k=-M}^{M} \mu_k z^k} \tag{5.4.1}$$

This scheme may be generalized by replacing the unit delay elements z^{-k} with *all-pass* sections of the form

$$\tilde{z}^{-1} = D_k(z^{-1}) = \left(\frac{z^{-1} - \lambda}{1 - \lambda \cdot z^{-1}}\right)^k \tag{5.4.2}$$

where λ is a *warping parameter* and $D_k(z^{-1})$ is a *warped delay element*. The phase function of a first-order all-pass filter $D_1(z^{-1})$is [Mat01]

$$\psi(\omega) = arg\left(D_1(e^{-j\omega})\right) = \tilde{\omega} = \omega + 2\arctan\frac{\lambda\sin\omega}{1 - \lambda\cos\omega} \tag{5.4.3}$$

This last equation is also known as the *frequency mapping function*. Therefore, the linear frequency axis ω is tranformed by (5.4.3) to a non-uniform frequency resolution (the warped frequency axis $\tilde{\omega}$), resulting in the frequency-warped spectrum $\tilde{S}(e^{j\tilde{\omega}})$.

Using a particular *warp factor* enables us to simulate the *Mel-frequency* or *Bark-frequency*

$$f_{\mathrm{Bark}} \triangleq 13 \cdot \arctan(0.00076 \cdot f_{\mathrm{Hz}}) + 3.5 \cdot \arctan((f_{\mathrm{Hz}}/7500)^2)$$

as shown in Figure 5.7. Furthermore, an adjustment of the warp factor can be used to achieve formant frequency shifts. In this way, the spectrum can be normalized to compensate for speaker-dependent differences in vocal tract length [Don00].

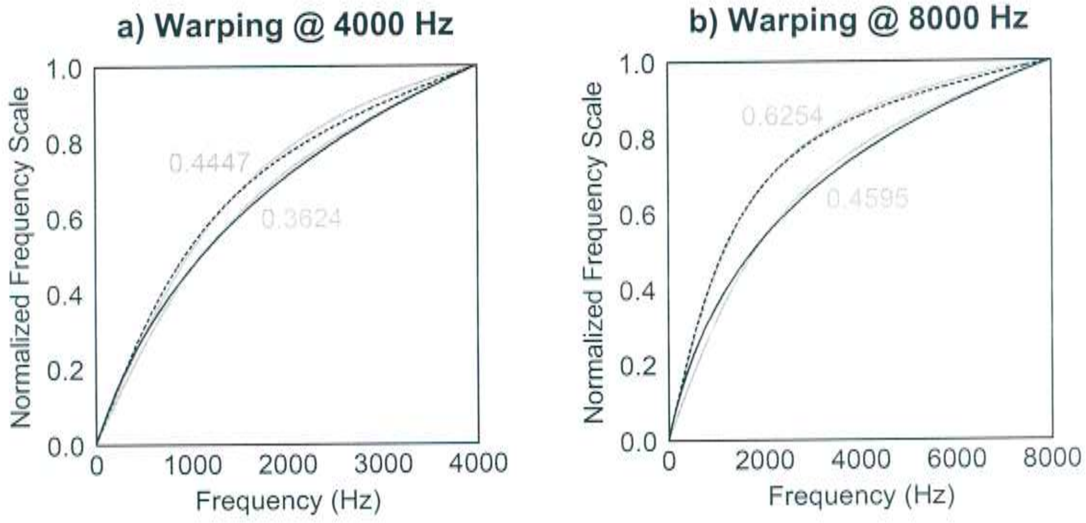


Figure 5.7: The frequency warp function in the bilinear transformation.

The prediction error minimization in the frequency-warped-domain is equivalent to minimize the mean square prediction error $\tilde{E}_{\mathrm{MV}}$ of the *inverse filter*

$$A_{\mathrm{MV}}(\tilde{z}(z)) = \sum_{k=-M}^{M} \tilde{\mu}_k \tilde{z}^{-n} \tag{5.4.4}$$

in the linear-frequency-domain. Given the relation to LP by the fast MVDR spectrum computation, the problem can be simplified to the minimization problem of the mean square prediction error $\tilde{E}_{\mathrm{LP}}$ of the *inverse filter*

$$\tilde{A}(z) = A(\tilde{z}(z)) = 1 + \sum_{n=1}^{N} \tilde{a}_n \tilde{z}^{-n}(z) \tag{5.4.5}$$

where $\tilde{a}_n$ denotes the n-th warped LPC (WLPC).

Warped LP (WLP) was first introduced by Strube [Str80] and systematically employed by other researchers [Tok95][Edl00], where in the latter a slightly different approach to WLP was introduced. To calculate the WLPC we have to apply the bilinear transformed signal to the autocorrelation method by the frequency-warped signal $\tilde{s}[n]$ which is defined by

$$\tilde{S}(z) = \sum_{n=0}^{\infty} \tilde{s}[n] \cdot \tilde{z}^{-n} = S(z) = \sum_{n=0}^{N-1} s[n] \cdot z^{-n} \qquad (5.4.6)$$

The inverse filter (5.4.5) on the Mel-frequency axis is therefore estimated by the Levinson-Durbin recursion, (4.1.16) to (4.1.18), replacing the short time autocorrelation coefficients $R(j)$, (**??**), by the *Mel-autocorrelation coefficients*:

$$\tilde{R}(j) = \sum_{m=1}^{\infty} \tilde{s}_n(m)\tilde{s}_n(m+j) \qquad (5.4.7)$$

As (5.4.6) shows, the bilinear transformation of a finite sequence results in an infinite sequence, and thus, the direct calculation of the Mel-autocorrelation coefficients (5.4.7) is not feasible. To overcome this problem different approaches were proposed [Str80]. In our work we want to follow an efficient implementation by Matsumoto et al. [Mat98] which is a Mel-autocorrelation method on the linear frequency axis.

The *total prediction error* on the Mel-frequency axis can be written in the linear frequency domain as

$$E_{\text{prediction}}(e^{-j\omega}) = \frac{1}{2\pi} \int_{-\pi}^{+\pi} \left| \tilde{A}(e^{-j\omega}) S(e^{-j\omega}) W(\omega) \right|^2 d\omega \qquad (5.4.8)$$

where

$$W(\omega) = \frac{\partial \psi(\omega)}{\partial \omega} = \frac{\sqrt{1-\lambda^2}}{1-\lambda e^{-j\omega}} \qquad (5.4.9)$$

Thus, the error minimization is equivalent to minimize the output error of $\tilde{A}(e^{-j\omega})$ excited by the pre-filtered signal

$$s_w(n) = W(\omega)s(n) \qquad (5.4.10)$$

Since $s_w(n)$ is an infinite sequence, and thus not tractable, Matsumoto et al. have proposed to remove $W(\omega)$. This is possible trough the introduction of the filter

$$\tilde{S}_w(z) = S(z)W^{-1}(z) \qquad (5.4.11)$$

resulting in a different inverse filter $\tilde{A}_w(z)$. Now, as a result of minimizing the total prediction error over the infinite time interval, the warped predictors can be obtained by solving

$$\sum_{j=1}^{p} \tilde{a}_{w,j}\phi(i,j) = -\phi(0,j) \quad : i = 1, \ldots, p \qquad (5.4.12)$$

where

$$\phi(i, j) = \sum_{n=0}^{\infty} y_i(n) y_i(n) \tag{5.4.13}$$

using the output sequence $y_i(n)$ of the i^{th} order all-pass filter exited by $y_0(n) = s(n)$. In terms of Parceval's theorem [Bro95] $\phi(i, j)$ is proved to be equal to the autocorrelation function $R_w(i-j)$ whose Fourier transform is equal to the warped and frequency-weighted PS $|\tilde{S}(e^{-j\tilde{\omega}})\tilde{W}(e^{-j\tilde{\omega}})|^2$. Therefore (5.4.12) becomes an autocorrelation equation as in conventional LP analysis. Since $\phi(i, j)$ is a function of the difference $|i-j|$, it can be calculated as the sum of finite terms

$$\phi(i, j) = \tilde{R}_w(|i-j|) = \sum_{n=0}^{N-1} s(n) y_{|i-j|}(n) \tag{5.4.14}$$

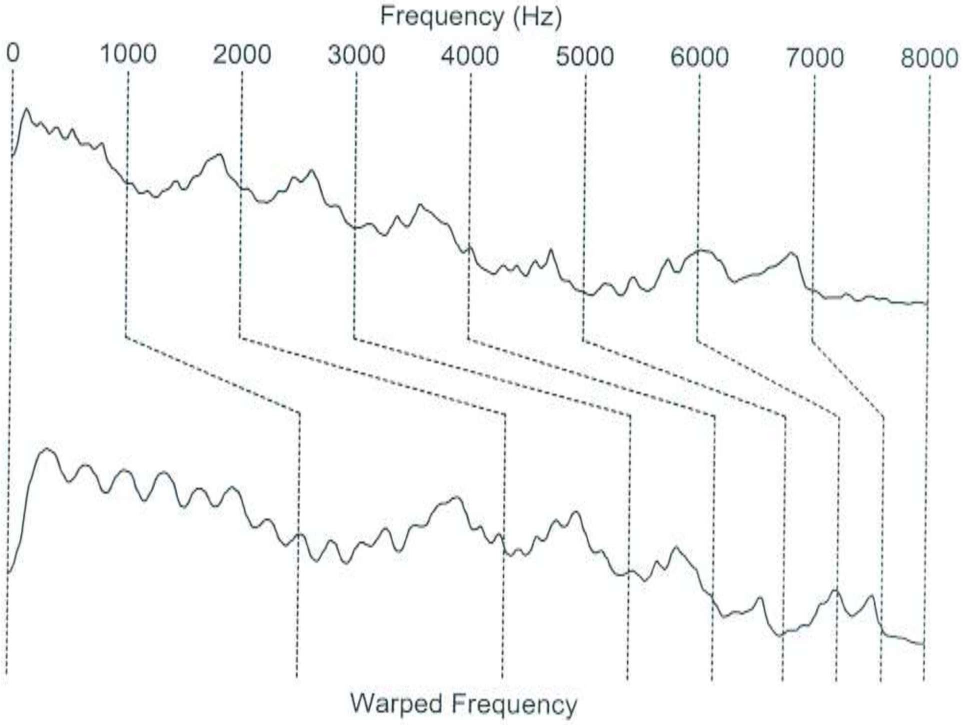


Figure 5.8: Comparison of unwarped and warped MVDR SE, both of order 120.

After obtaining the prediction coefficients $\tilde{a}_{w,i}$ by solving the Levinson-Durbin recursion, (4.1.16) to (4.1.18), with $R(j)$, (5.4.14) instead of (??) the WLPC can be easily obtained by

$$\tilde{a}_i = \lambda_0 \tilde{a}_{w,i} + \lambda_1 \left(\tilde{a}_{w,i-1} + \tilde{a}_{w,i+1} \right) \tag{5.4.15}$$

where

$$\lambda_0 = (1+\lambda^2)(1-\lambda^2)^{-1/2} \tag{5.4.16}$$
$$\lambda_1 = \lambda(1-\lambda^2)^{-1/2} \tag{5.4.17}$$

As already mentioned before, these WLPC can now be used to calculate the WMVDR by the fast MVDR spectrum computation.

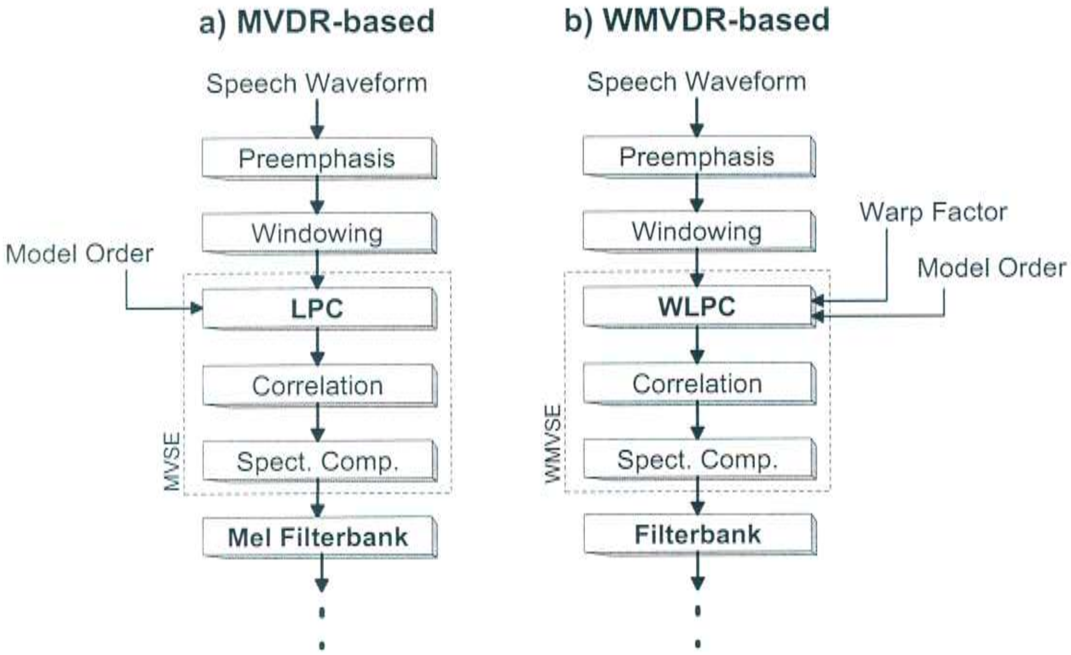


Figure 5.9: Extract of the MVDR and the WMVDR frontend.

Figure 5.8 illustrates the ordinary and warped MVDR SE. While the MVDR exhibits frequency-independent inherent spectral resolution, the warped MVDR (WMVDR) provides a high resolution on frequencies below 2000 Hz (warp factor 0.4595 ~ Mel) with decreasing resolution to higher frequencies. The warping of the MVDR provides an interesting property which can't be established equally by a MVDR followed by a frequency-warping:

Similar to WLP [Kar01] the WMVDR residuals show spectral flattering and level compensation similar to the adaptation of the firing rate in the auditory nerve, resulting in information of the residuals which resembles the overall information in the auditory nerve firing.

Figure 5.9 shows an extract of the MVDR and the WMVDR frontend. Note that the Mel-frequency warping is achieved differently.

5.5 MVDR Spectral Estimation in Noise

The effect of additive noise on the SE broadens the spectral peaks and displaces them from their true positions. Compare a) with b) and c), and d) with e) and f) of Figure 5.10. In other words, the spectral resolution decreases as the SNR decreases [Kay88]. This effect can be explained by the fact that the all-pole assumption is no longer valid if additive noise is present. Under the assumption of uncorrelated white noise of variance σ_w^2, the PSD can be written as

$$S_{xx}(z) = \frac{\sigma^2}{A(z)A^*(1/z^*)} + \sigma_w^2 = \frac{\sigma^2 + \sigma_w^2 A(z)A^*(1/z^*)}{A(z)A^*(1/z^*)} \tag{5.5.1}$$

Thus, the PSD is characterized by zeros as well as poles while the dynamic range is reduced. Since the prediction error filter $A(z)$ attempts to whiten the PSD, the zeros of $A(z)$ are located near the unity circle for low SNR, because the PSD is already flat due to noise. Thus the subsequent filtering will not whiten the PSD further. It follows that the AR spectral estimate in the presence of noise is a smoothed version of the SE which would have been obtained if no noise was present. For a detailed discussion, see [Kay88].

Note that due to additive noise the various methods derived on the ML principles are no longer ML solutions. Instead for large data records a function given by Hosoya [Hos79]

$$\int_{-1/2}^{+1/2} \left[\ln S_{xx}(f) + \frac{S_{\mathrm{PER}}(f)}{\ln S_{xx}(f)} \right] df \tag{5.5.2}$$

where

$$S_{xx}(f) = \frac{\sigma^2}{|A(f)|^2} + \sigma_w^2 \tag{5.5.3}$$

and

$$S_{\mathrm{PER}}(f) = \frac{1}{N}\left|\sum_{n=0}^{N-1} x[n]e^{-j2\pi fn}\right|^2 \tag{5.5.4}$$

has to be minimized. This is a difficult nonlinear minimization problem. To our knowledge, only two solutions have so far been proposed, both of which are based on the Newton-Raphson approach and suffer from the usual problems of local minima and lack of convergence.

To deal with this, some suboptimal estimators have been proposed. They involve:

- **Use of pole-zero estimators**
  This approach recognizes that the noisy model becomes a pole-zero model. But, as we have seen, the all-pole model has the advantage of modelling the perceptually important spectral peaks better than the spectral valleys. Hence, the pole-zero approach is not very useful for present purposes.

- **Pre-filtering**
  Before an all-pole estimator is applied, the signal is enhanced by a Wiener filter which can be shown to be an implicit part of the ML estimation. This approach is quite successful for speech data [Lim78].

- **all-pole parameter compensation (noise compensation)**
  In this method, the bias due to additive noise is removed. A drawback of this method is that it results in overly peaky estimates. Moreover, the autocorrelation matrix (4.1.15) may become ill conditioned leading to a spectral estimator with large variance [Kay88].

- **High order all-pole modelling**
  To reduce the bias due to the model mismatch, a high order may be used. The feasibility is guaranteed by the Kolmogorov theorem [Pri81], which states that an all-pole model of infinite order adequately models any wide sense stationary process. The shortcoming of this approach is that, as we have already mentioned, spurious peaks may appear with increasing model order.

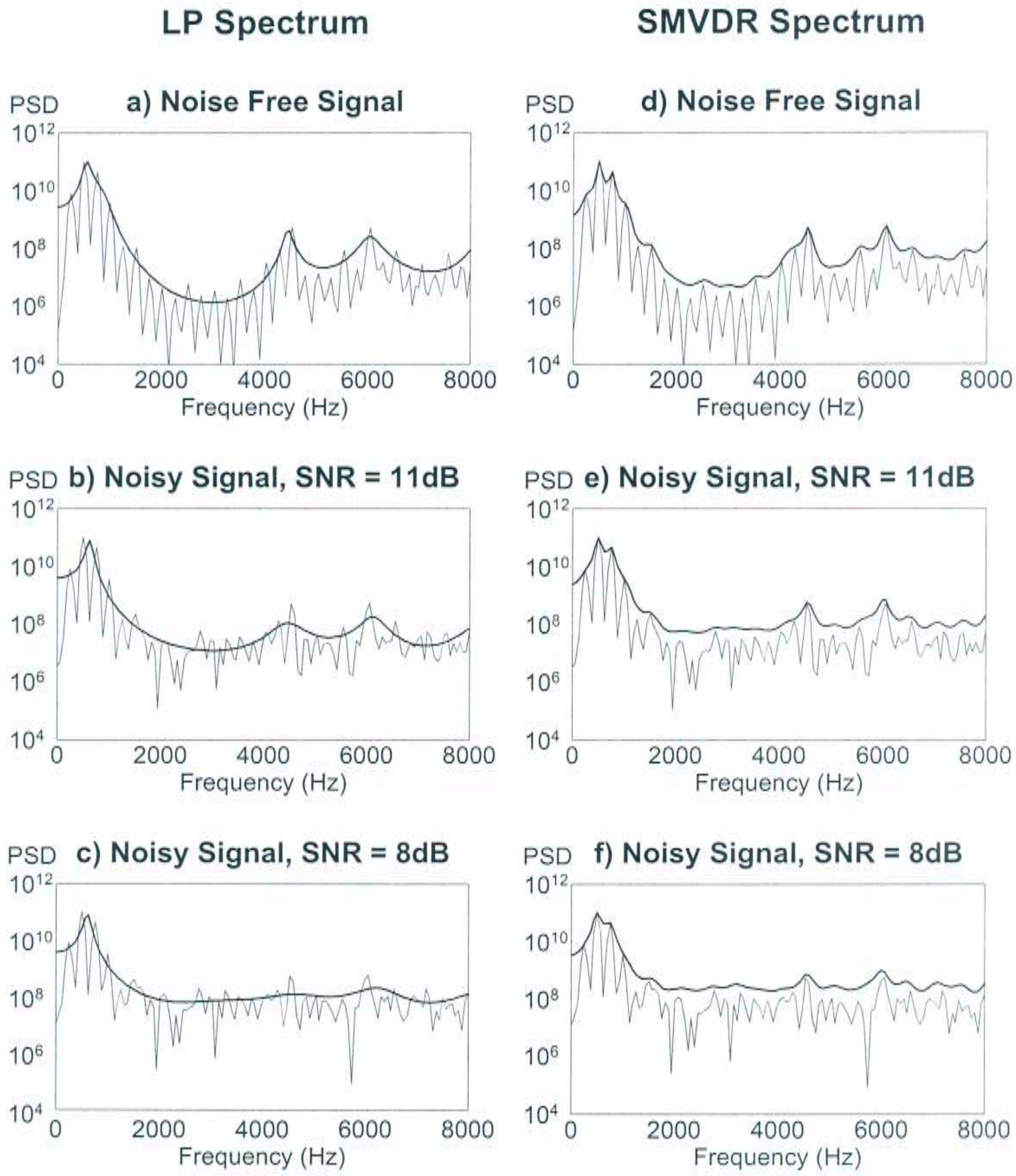


Figure 5.10: Comparison of the behavior of FFT (thin line), LP of order 20 and SMVDR of order 100 in different noise levels.

5.6 Comparisons of MVDR and LP

Before closing the present chapter, we develop several important relations between MVDR and LP spectral estimates. An understanding of these relations leads to useful insights.

Recalling that the Cholesky decomposition of ϕ_{xx}^{-1} is

$$\phi_{xx}^{-1} = \mathbf{B}\mathbf{P}^{-1}\mathbf{B}^H \tag{5.6.1}$$

we can analytically relate the LP spectral estimator to the MVDR spectral estimate [Bur72]. Substituting (5.6.1) into (5.1.18) provides

$$S_{\mathrm{MV}}(\omega) = \frac{1}{\mathbf{s}^H(\omega)\mathbf{B}\mathbf{P}^{-1}\mathbf{B}^H\mathbf{s}(\omega)} \tag{5.6.2}$$

with the relation of

$$\begin{aligned}
\mathbf{B}^H\mathbf{s} &= \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ a_1^{(1)} & 1 & 0 & \dots & 0 \\ a_2^{(2)} & a_2^{(1)} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_p^{(p)} & a_{p-1}^{(p)} & a_{p-2}^{(p)} & \dots & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ e^{j\omega} \\ e^{j2\omega} \\ \vdots \\ e^{jp\omega} \end{bmatrix} \\
&= \begin{bmatrix} 1 \\ a_1^{(1)} + e^{j\omega} \\ a_2^{(2)} + a_1^{(2)}e^{j\omega} + e^{j2\omega} \\ \vdots \\ a_p^{(p)} + a_{p-1}^{(p)}e^{j\omega} + \dots + e^{jp\omega} \end{bmatrix} \qquad (5.6.3) \\
&= \begin{bmatrix} 1 \\ e^{j\omega}\mathbf{a}^{(1)}(\omega) \\ e^{j2\omega}\mathbf{a}^{(2)}(\omega) \\ \vdots \\ e^{jp\omega}\mathbf{a}^{(p)}(\omega) \end{bmatrix}
\end{aligned}$$

where

$$\mathbf{a}^{(k)}(\omega) = 1 + \sum_{i=1}^{k} a_i^{(k)} e^{-ji\omega} \tag{5.6.4}$$

we may rewrite (5.6.2) as

$$S_{\mathrm{MV}}(\omega) = \frac{1}{\sum_{k=0}^{p} |\mathbf{B}^H\mathbf{s}|_k^2/\rho_k} \tag{5.6.5}$$

if ρ_k denotes the prediction error power for the k^{th} order predictor and the elements of $\mathbf{B}^H\mathbf{s}$ are indexed $k = 0 \dots p$. Rewriting the denominator we yield

$$\frac{|\mathbf{B}^H\mathbf{s}|^2}{\rho_k} = \frac{|\mathbf{a}^{(k)}(\omega)|^2}{\rho_k} = \frac{1}{S_{\mathrm{LP}}^{(k)}(\omega)} \tag{5.6.6}$$

where $S_{\mathrm{LP}}^{(k)}(\omega)$ is the k-th order LP spectrum. Now we can write the *relationship between MVDR and LP spectra*:

$$\frac{1}{S_{\mathrm{MV}}^{(p)}}(\omega) = \sum_{k=0}^{p} \frac{1}{S_{\mathrm{LP}}^{(k)}(\omega)} \tag{5.6.7}$$

This implies that the MVDR spectrum $S_{\mathrm{MV}}^{(p)}(\omega)$ of p^{th} order is the harmonic mean of the LP spectra of order 0 to p. This relationship is also a good explanation why, in general, the MVDR spectrum exhibits a smoother frequency response with decreased variance than the corresponding LP spectrum [Mur00].

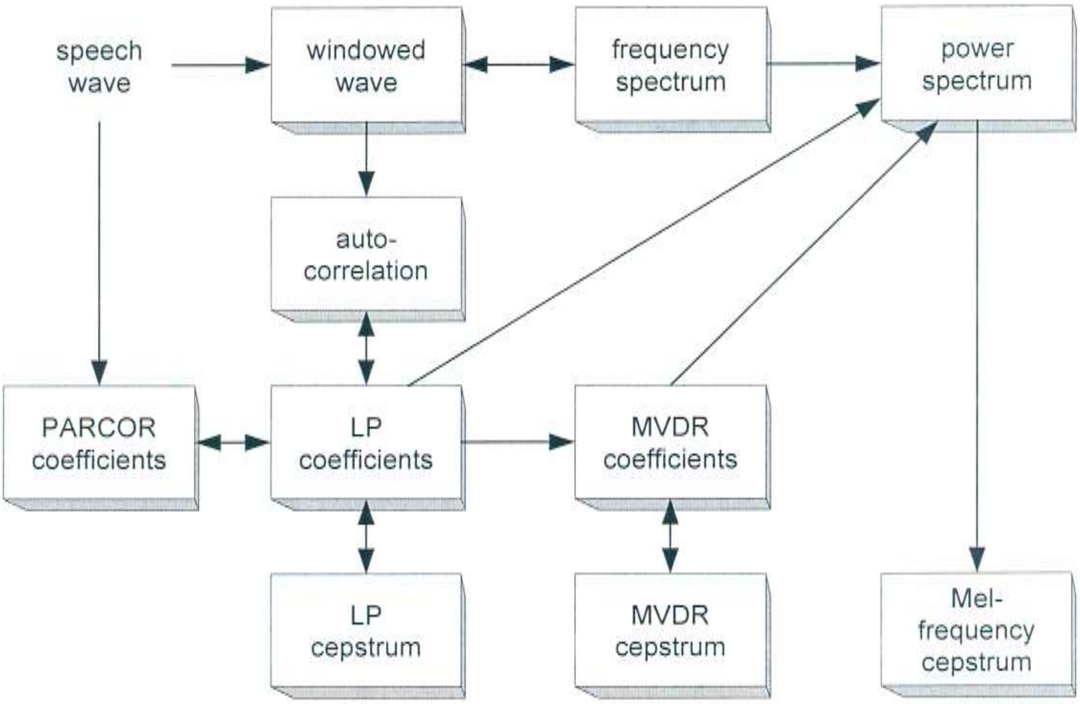


Figure 5.11: Mutual relationships between spectrum, coefficients and cepstrum

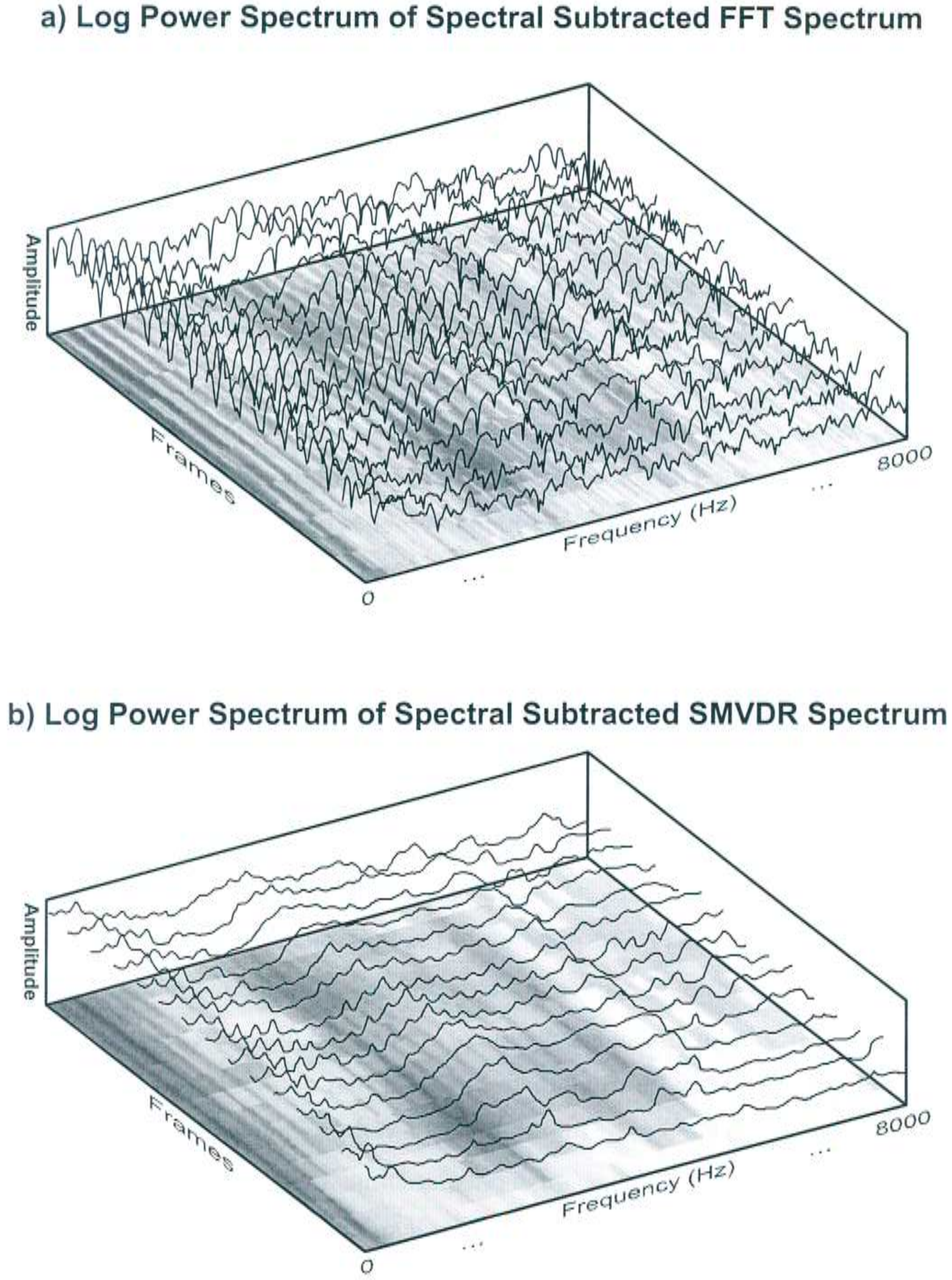
Figure 5.11 shows the relationships between different representations of speech parameters. An arrow with one end symbolizes a transformation which is not invertible while an arrow with two ends symbolizes a transformation with an existing inverse. Note that even if two or more arrows end in the same knot, it doesn't mean that the parameters are the same.

6 Noise Estimation and Subtraction

The underlying principle of *spectral subtraction* (SS) is the subtraction of the noise from the noise-contaminated signal in the spectral domain. The prerequisite for this technique, is that the noise spectrum is known or can be estimated. In some applications, it is possible to use two microphones, the first of which receives the noise corrupted speech from the speaker, the second of which is positioned so as to receive only noise. In this case the noise spectrum is obtained directly from the second microphone. In applications where the use of two microphones is impossible, whole utterances can be segmented into speech and non-speech regions, and the noise spectrum can be estimated from those regions containing no speech.

SS has proven useful in increasing the perceptual quality of speech signals corrupted by additive noise. The main drawback of this technique is that the noise remaining after the processing has a very unnatural quality [Bol79, Cap94]. This can be explained by the fact that the magnitude of the short-time PS exhibits strong fluctuations in noisy areas. After the spectral attenuation, the frequency bands that originally contained the noise consist of randomly spaced spectral peaks corresponding to the maxima of the short-time PS. Between these peaks, the short-time PS values are close to or below the estimated averaged noise spectrum, which results in strong attenuations. As a result, the residual noise is composed of sinusoidal components with random frequencies that come and go in each short-time frame [Bol79]; e.g., see spectra of Figure 6.1 a) and the arrow in Figure 6.2 b). These artifacts are known as *musical*[1] *tones/noise* phenomenon. One way to reduce this unwanted effect is to median smooth the signal after spectral subtraction. Unfortunately, this leads to audible signal distortions [Lin97]. To overcome this problem, we propose the use of a high resolution SE, such as that provided by SMVDR, instead of smoothing. As we will show, cascading spectral subtraction and SMVDR spectral estimation provides for effective noise

[1] This term is a reference to the presence of pure tones in the residual noise.



a) Log Power Spectrum of Spectral Subtracted FFT Spectrum

b) Log Power Spectrum of Spectral Subtracted SMVDR Spectrum

Figure 6.1: Logarithmic Power Spectra of Spectral Subtracted FFT and SMVDR Spectrum.

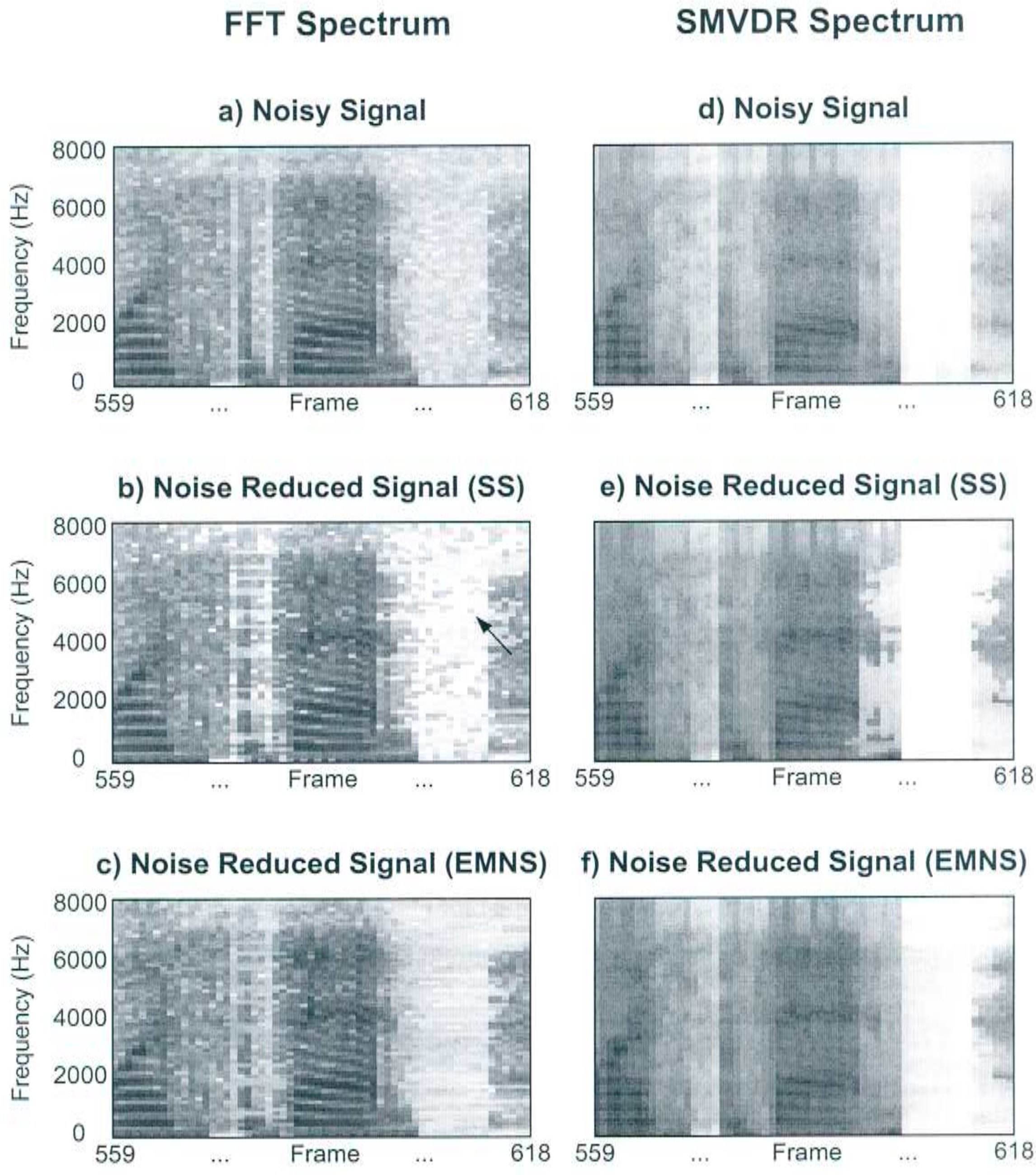


Figure 6.2: Spectral subtraction applied on FFT- and SMVDR spectra
(EMNS = Ephraim and Malah noise suppressor)

reduction, while simultaneously minimizing the introduction of musical tones and other audible signal distortions.

6.1 Noise Estimation

Noise estimation is not our primary concern here. The noise estimator, however, has a major impact on the overall quality of a SS system, especially if the algorithm should be capable of handling nonstationary noise. Hence, the noise estimation method used in our evaluations will be described here briefly.

Speech enhancement based on *minimum statistics* was proposed in [Mar94] and modified in [Dob95]. In contrast to other methods, the minimum statistics algorithm does not use any explicit threshold to distinguish between speech activity and speech pauses and is therefore more closely related to soft-decision methods than to the traditional voice activity detection methods. As in other soft-decision methods, it can update the estimated noise PSD during speech activity. It was recently confirmed [Mey97] that the minimum statistics algorithm performs well in nonstationary noise.

The minimum statistics method rests on two observations, namely that the speech and the disturbing noise are statistically independent and that the power of a noisy speech signal frequently decays to the power level of the disturbing noise. It is therefore possible to derive an accurate noise PSD estimate by tracking the minimum of the noisy signal PSD. Since the minimum is smaller than the average value, the minimum tracking method requires a bias compensation. A detailed description about minimum statistics can be found in [Mar01].

6.1.1 The Advantage of Spectral Envelopes over Fourier Spectra for Noise Estimation

Comparing frame based autocorrelations over different spectra (Fourier, MVDR and SMVDR) of different speech- and noise environments of Figure 6.3, it is immediately apparent that the frames of MVDR and SMVDR spectra are more strongly correlated than the frames of the Fourier spectrum. An examination of the LP spectrum leads to a similar conclusion, and hence is not shown here for reasons of conciseness. Since noise estimation is dependent on the assumption of short-term stationarity, any improvement in stationarity,

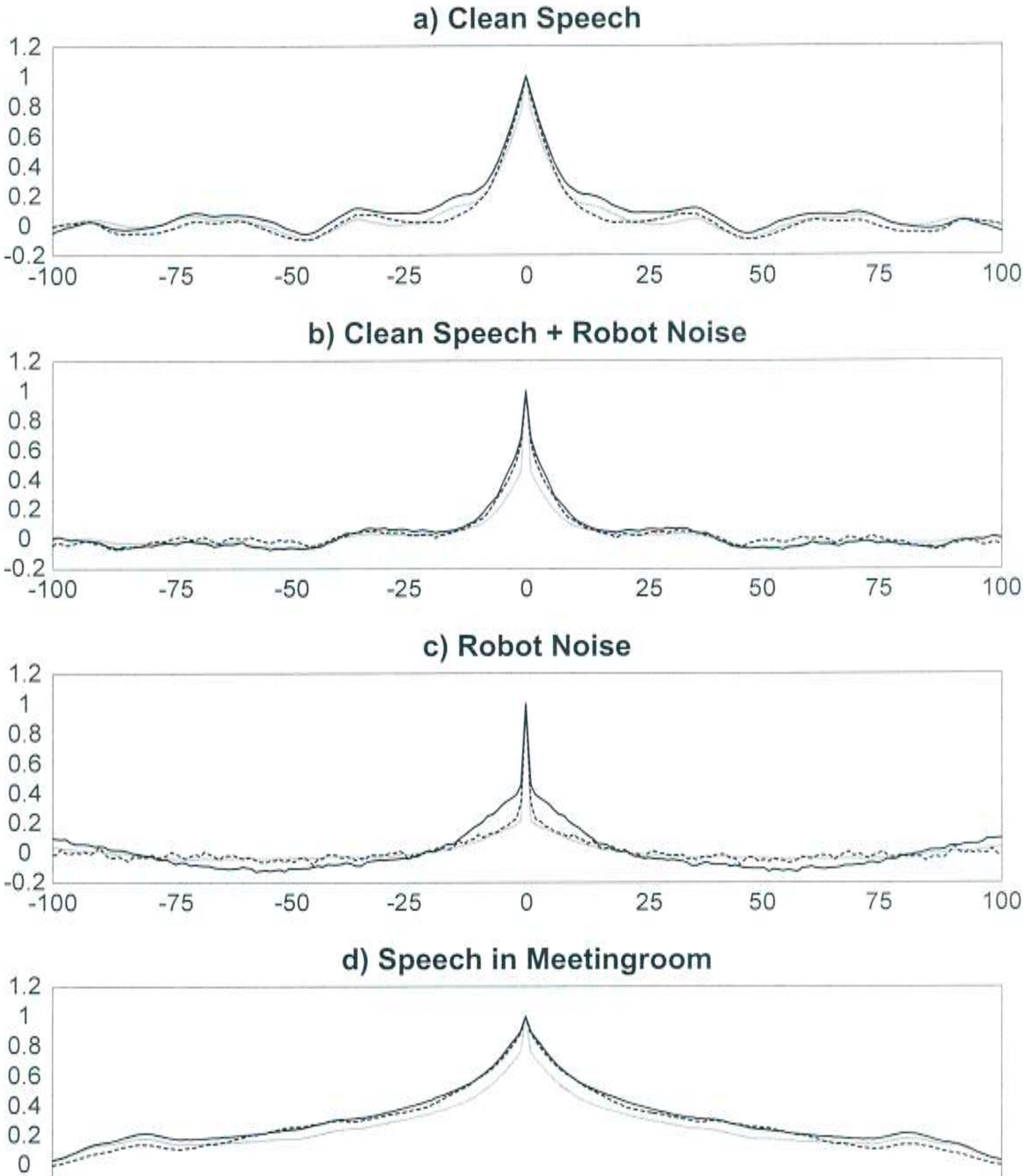


Figure 6.3: Frame based autocorrelation (mean subtracted and averaged over all frequencies), black: SMVDR Spectrum, pointed: MVDR Spectrum, gray: FFT Spectrum

measurable through an improvement in autocorrelation, results in a better approximation of the noise. This in turn leads to a further improvement of the reconstructed signal with fewer musical tones.

6.2 Spectral Subtraction

Assuming that the clean speech signal $s(t)$ is corrupted by uncorrelated additive noise $n(t)$, the disturbed signal can be written as:

$$y(t) = s(t) + n(t) \tag{6.2.1}$$

This implies that the PSD can be expressed as

$$Y(\omega_k) = S(\omega_k) + N(\omega_k) \tag{6.2.2}$$

Thus, to get an estimate of the undisturbed signal $\hat{S}(\omega_k)$, we must subtract the estimated noise power $\hat{N}(\omega_k)$, as estimated from regions without speech activity, from the noisy speech power at every frequency band.

$$\hat{X}(\omega_k) = Y(\omega_k) - \hat{N}(\omega_k) \tag{6.2.3}$$

Thus the name *spectral subtraction*. Since the estimated noise power $\hat{N}(\omega_k)$ can be larger than the disturbed signal power, thereby resulting in a negative signal power, equation (6.2.3) must be modified to suppress negative power:

$$\hat{X}(\omega_k) = \max\left\{Y(\omega_k) - \hat{N}(\omega_k), \alpha \geq 0\right\} \tag{6.2.4}$$

In the above, the parameter α represents the *spectral floor*.

The use of noise subtraction for ASR has proven more successful than it has for speech enhancement [Mor88]. This is due to its use of spectral representation as the feature space which makes it unnecessary to recreate the speech signal requiring a phase information which is commonly approximated by the phase information given by the disturbed signal.

Different modifications of the basic SS have been proposed to reduce musical tones [Bol79, Vas92], but all have failed to completely eliminate them. Ephraim and Malah [Eph84] have been proposed a noise suppression technique which is able to avoid musical tones while still obtaining a significant noise suppression; see Figure 6.2 c).

6.2.1 The Advantage of Spectral Subtraction over Smoothed Spectra for Noise Subtraction

Smoothing a spectrum by applying a low-pass filter $\overline{S} = S * F_{\text{lowpass}}$ results in a higher correlation ρ of neighbored frequencies under the drawback of information loss (measured in entropy)

$$-\sum_{\overline{x}\in\overline{\chi}} p(\overline{x}) \log p(\hat{x}) \leq -\sum_{x\in\chi} p(x) \log p(x) \quad : \overline{\chi} \leq \chi \tag{6.2.5}$$

because the size of the frequency-band and thus the number of possibilities are reduced.

A higher correlation between neighboring frequency spectra result in fewer musical tones [Bol79] which increase the accuracy of a speech recognition system, while the loss of information decreases the accuracy. Therefore, we are confronted with a tradeoff; we seek a feature which is able to increase the correlation of neighboring frequencies without loosing *relevant* information. As hearing or, more precisely spoken word understanding, is not fully understood, it is difficult to quantify what information is relevant. Nevertheless, in the speech recognition community it is widely believed that most of the relevant information lies in the spectral peaks. Therefore we already have investigated a feature providing a high correlation of neighbored frequencies while remaining a high grade of information; to wit, the spectral estimate derived from the MVDR or SMVDR.

6.2.2 Influence of Spectral Subtraction on the Log-Spectral Features

Figure 6.4 a) clearly demonstrates the ability of SS to reduce the negative effect of additive noise on the features. Its drawback is a splitting of the features into two conurbations resulting in attenuations of the spectra by randomly spaced spectral peaks. Figure 6.4 b) demonstrate that using the Ephraim and Malah noise suppressor (EMNS), instead of SS, is indeed able to overcome the problem of conurbation separation, but with the obvious drawback of a limited reduction of feature distortion.

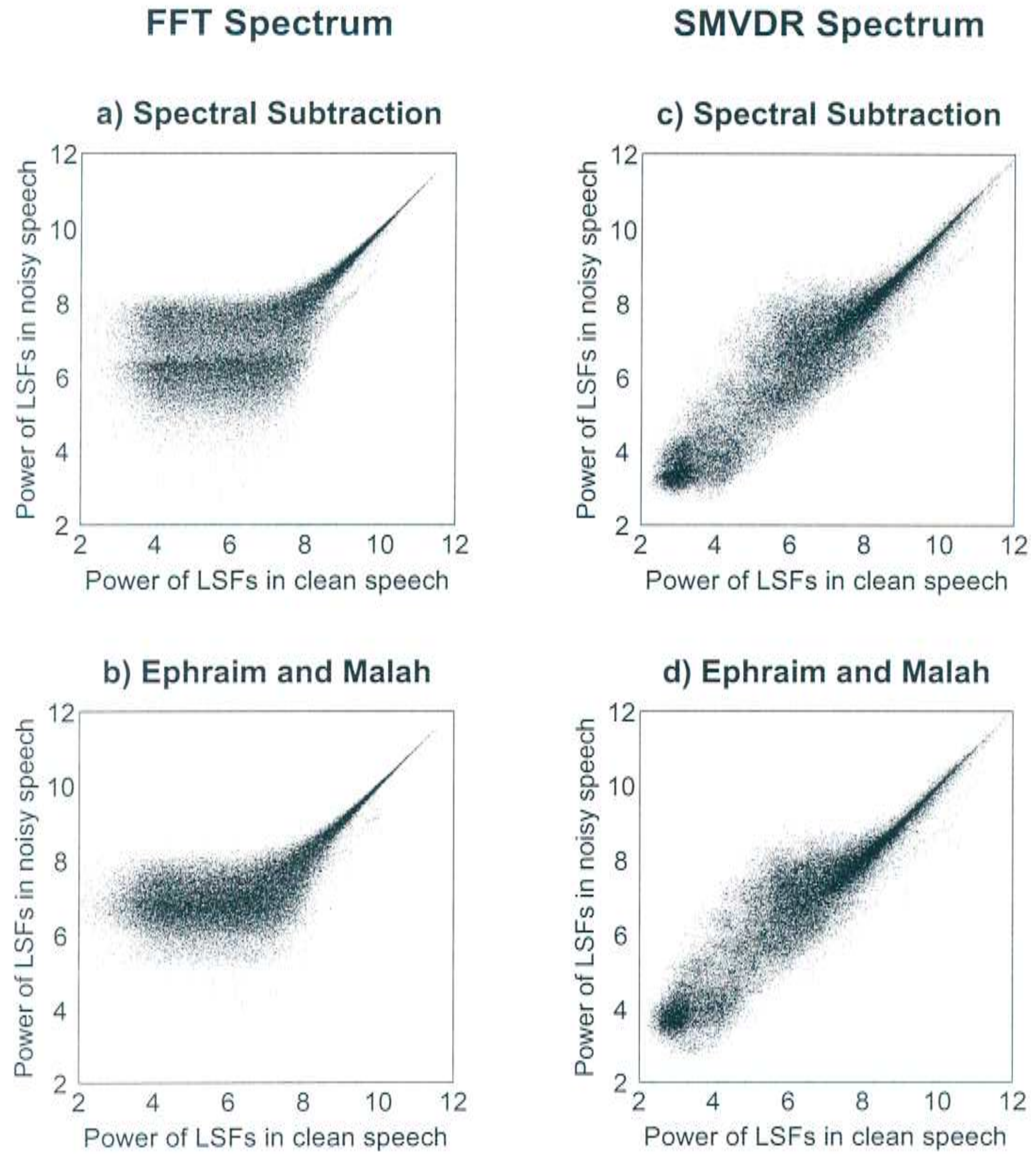


Figure 6.4: Comparison of log-spectral features in the power-spectral domain.

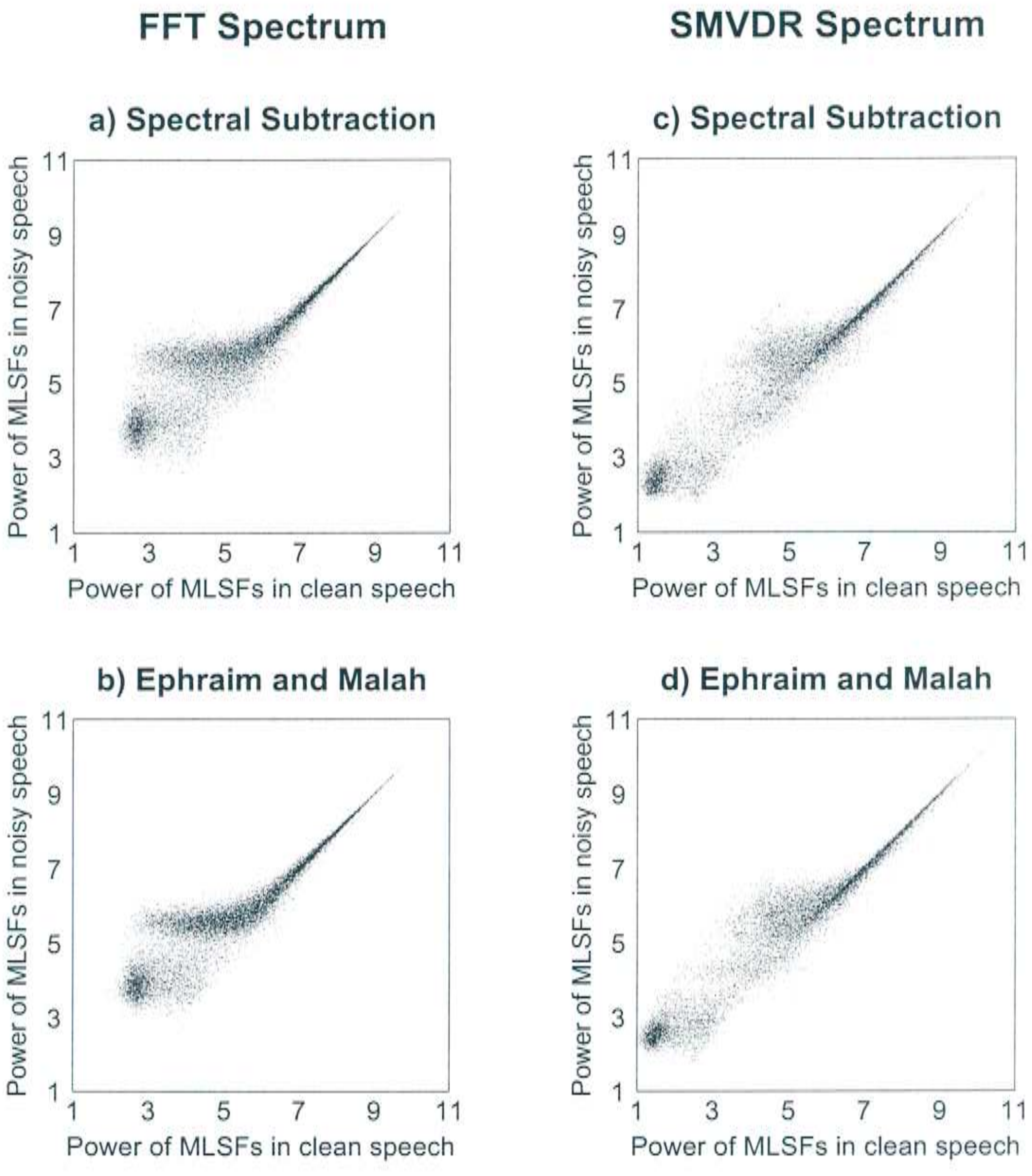


Figure 6.5: Comparison of log-spectral features in the Mel-spectral domain.

As seen in Chapter 5.3, the SMVDR based features are less distorted by additive noise than the Fourier based features. A further reduction of feature distortion is possible by applying SS on the SMVDR based features. Here a separation into different conurbations does not occur, thereafter no randomly spaced spectral peaks are occurring which makes SS based on the SMVDR envelope a suitable method to reduce feature distortion without the appearance of musical tones.

7 Speech Recognition Experiments

In this chapter we present the results of several speech recognition experiments undertaken to determine the effectiveness of spectral estimation based on the MVDR and their variations.

7.1 Training and Test Conditions

Recognition performance was tested in different acoustic conditions using different degrees of *mismatch* between training and test. The term *matched conditions* refers to an experimental set-up wherein the acoustic conditions during training and recognition were identical. The term *mismatched conditions*, on the other hand, refers to experiments wherein the acoustic models were trained on clean data, but recognition was performed on noisy data. Recognition performance was also determined for an *intermediate condition*, where the acoustic models were trained on noisy data that matched the noise at recognition time. Two distinct spontaneous speech corpora were used for the experiments described below.

Switchboard Corpus

For the first set of speech recognition experiments, training and testing was conducted on the *Switchboard Corpus*. Switchboard consists of spontaneous conversations collected over standard American T1 telephone lines. Switchboard contains approximately 240 hours of speech and about 3 million words of text, contributed by 500 speakers of both sexes from every major dialect of American English. The total corpus is divided into 2,430 conversations, each averaging 6 minutes in length. To reduce the experimental turnaround time, we used a gender balanced subset of approximately 31 hours of speech for training. Our test set contained 8,186 words of speech contributed by 16 speakers of both sexes.

For the Switchboard experiments, we used a baseline model with 32 Gaussians for each of

4,166 codebooks for a total of 133,312 Gaussians. All features were calculated every 10 ms from speech data sampled at 8 kHz, using either a 20 ms Hamming window for FFT and autocorrelation based methods, or a rectangular window for Burg based method. FFT-, LP-, MVDR- and SMVDR-based MFCC were used followed by a speaker-dependent frequency domain VTLN. Thirteen cepstral components, along with their first and second differences were derived using a Mel-Filterbank with 30 half-overlapping triangularly shaped filters followed by a DCT. LDA was used to reduce the final feature length to 32.

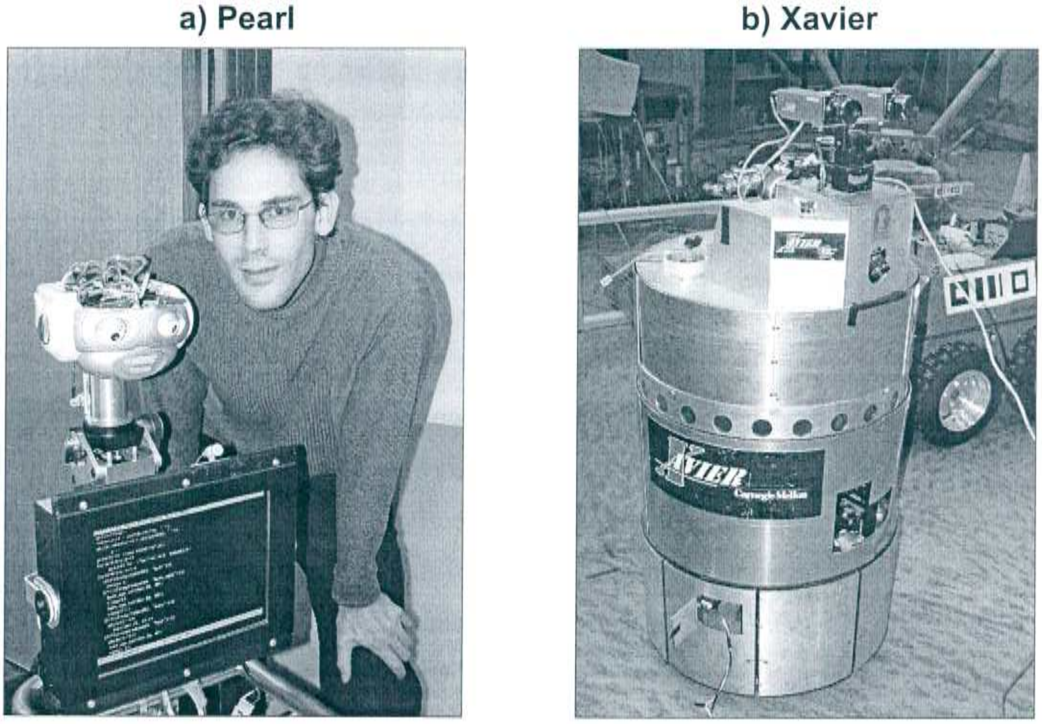
English Spontaneous Scheduling Task (ESST)

The second set of speech recognition experiments was conducted on speech material from the *English Spontaneous Scheduling Task* (ESST) corpus, which consists of spontaneous speech collected during dialogues between two persons who are making business travel arrangements. This speech was recorded with Sennheiser close-talking microphones in a noise-free environment. Testing was done using the original clean speech, as well as clean speech plus additive noise. Additional testing was conducted on the ESST data in the presence of reverberant distortion, as described below. The ESST training set contains approximately 35 hours of speech contributed by 242 speakers of both sexes. The test set contains 22,899 words spoken by 15 speakers of both genders.

For these experiments, we used a baseline model with 48 Gaussians for each of 2,339 codebooks, giving a total of 112,272 Gaussians. All features were calculated every 10 ms from speech data sampled at 16 kHz, using either a 20 ms sliding Hamming window for FFT and autocorrelation based methods, or rectangular window for Burg based methods. Subsequent processing was identical to that described above.

Noise and Reverberation

To test the proposed spectral estimation techniques under mismatched and intermediate conditions, several different types of noise were used. The first of these was simple *white noise* while the next type of noise was collected from a variety of *humanoid robots*; robot noise has, in turn, a number of distinct sources: *Background noise*, says Yoshihiro Kuroki [Wil02], a senior manager at the Digital Creators Laboratory (Sony Humanoid Robot project) "is proving particularly difficult for the engineers to tackle." Also additional distortions may arise when, for example, two or more people speak simultaneously: The *attention tracking* problem — whom to listen to and the so-called *cocktail party effect* — if noise appears, people change there style of speaking.



a) Pearl

b) Xavier

Figure 7.1: Photo of the robots used for noise recording.

For our experiments two robots where used to record noise. A brief description of the robots, see Figure 7.1, is given below:

- *Pearl* is a Nursebot who has two primary functions:

  1. reminding people about routine activities such as eating, drinking, taking medicine, and using the bathroom, and
  2. guiding them through their environments.

  It is equipped with a differential drive system, two on-board Pentium PCs, wireless Ethernet, SICK laser range finders, sonar sensors, microphones for speech recognition, speakers for speech synthesis, a touch-sensitive graphical display, a actuated head unit, and a stereo camera system.

- *Xavier* was designed to provide a test-bed for research.

  The base is a 24 zoll diameter, four-wheeled, synchro-drive mechanism. Sensors include bump panels, a Denning sonar ring, Nomadics laser light striper, and a Sony color camera mounted on a Directed Perception pan/tilt head. On-board computation consists of two 66 MHz Intel 486 computers, and an on-board color 486 laptop, all connected to each other via Ethernet and connected to the outside world via a Wavelan wireless card. Communication with Xavier is graphical, remote, and speech-driven. An off-board Next computer runs the Sphinx real-time, speaker independent ASR system and a text-to-speech board provides speech generation.

Whenever speech is recorded in an enclosed space using microphones in the medium or far fields, the resulting speech is distorted by *reverberation*. This is so because the final signal contains not only the speech emanating directly from the speaker's mouth to the microphone, but also delayed and filtered versions of the same which are caused by multiple reflections from walls and other sufaces. Since reverberation has a long duration in comparison to the typical speech analysis frame, this distortion manifests itself as a temporal smearing of the short-term power spectrum typically used as the basis for speech recognition features [Gel02]. In order to determine the effectiveness of MVDR spectral estimation for this type of distorted speech, the ESST data was played through a loud speaker into a meeting room, and subsequently recorded with an array of microphones located approximately two meters from the loudspeaker. The 16 channels of the microphone array were subsequently combined using the simple *delay and sum* algorithm.

7.2 Janus Recognition Toolkit

All speech recognition experiments presented in this thesis were conducted with the *Janus Recognition Toolkit* (JRTk), which is developed and maintained jointly at the Universität Karlsruhe (TH), Germany and at the Carnegie Mellon University in Pittsburgh, Pennsylvania, USA. Janus provides a flexible Tcl/Tk scripting environment, which allows for the rapid development of state-of-the-art speech recognizers.

7.3 Experimental Results

Here we present and discuss the results of our initial speech recognition experiments.

Switchboard Experiments

Our first set of experiments were conducted on the Switchboard Corpus. Due to the distortions introduced when speech is transmitted over T1 telephone lines, the results can be viewed as falling under the intermediate condition. The absolute and relative word error rates given in Tables 7.1 and 7.2 respectively show that using the right model order (e.g., 80) and LPC type (i.e., autocorrelation), the SMVDR based approach is able to outperform the Fourier based approach. A clear superiority of the MVDR based approaches over the Fourier based approaches, is not evident. In particular the reduction in WER is dependent on the model order and shows no clear trend; i.e., the SMVDR based speech recognizer of model order 80 and 120 performs better than the Fourier based speech recognizer, but in the case of model order 100 its performance is worse. This is not surprising given that the distortion introduced by telephone lines is not primarily additive in nature. Indeed, using SS on the SMVDR based MFCC using the autocorrelation method, order 80, increased the WER from 35.9% to 37.9% .

type	speaker	FFT-b. MFCC	MVDR-b. MFCC	SMVDR-based MFCC			
order			60	60	80	100	120
	average	36.2	*	*	*	*	*
	mean	37.2	*	*	*	*	*
autocorr.	average	*	36.8	36.2	35.9	36.5	35.9
	mean	*	37.7	37.2	36.7	37.4	36.8
Burg	average	*	36.4	36.8	36.8	36.9	37.5
	mean	*	37.4	37.9	37.8	37.9	38.5

Table 7.1: Comparison of word error rates (in percentage) on telephone speech.

In Appendix B the WER and RER are given for every speaker of the test set.

type	speaker	MVDR-b. MFCC	SMVDR-based MFCC			
order		60	60	80	100	120
autocorr.	average	- 1.7	+ 0.0	+ 0.8	- 0.8	+ 0.8
	mean	- 1.3	+ 0.0	+ 1.3	- 0.5	+ 1.1
Burg	average	- 0.6	- 1.7	- 1.7	- 1.9	- 3.6
	mean	- 0.5	- 1.9	- 1.6	- 1.9	- 3.5

Table 7.2: Comparison of relative error reductions (in percentage) on telephone speech.

speaker	SMVDR-based MFCC, autocorrelation, order 80			
smooth	1	2	3	4
average	36.3	35.9	36.0	36.1
mean	37.2	36.7	36.8	37.1

Table 7.3: Comparison of word error rates (in percentage) on telephone speech using different smoothing factors.

speaker	SMVDR-based MFCC, autocorrelation, order 80			
smooth	1	2	3	4
average	- 0.3	+ 0.8	+ 0.6	+ 0.3
mean	+ 0.0	+ 1.3	+ 1.1	+ 0.3

Table 7.4: Comparison of relative error reductions (in percentage) on telephone speech using different smoothing values.

Tables 7.3 and 7.4 show the effect of different smoothing factor using SMVDR-based MFCC, autocorrelation of order 80.

Experiments with Warped Envelopes on Telephone Speech

The speech recognition experiments conducted in this section where similar to the former, except that no VTLN was conducted. The results given in Tables 7.5 and 7.6 show that using the right model order (50) of the warped spectrum is superior to the Fourier and SMVDR based approaches.

speaker	FFT-b. MFCC	SMVDR-b. MFCC	SWMVDR-based MFCC					
order		80	40	50	60	70	80	90
average	39.1	38.7	38.7	38.4	38.7	38.5	38.7	38.5
mean	40.0	39.8	39.9	39.5	39.7	39.4	39.7	39.4

Table 7.5: Comparison of word error rates (in percentage) on telephone speech.

speaker	SMVDR-b. MFCC	SWMVDR-based MFCC					
order	80	40	50	60	70	80	90
average	+ 1.0	+ 1.0	+ 1.8	+ 1.0	+ 1.5	+ 1.0	+ 1.5
mean	+ 0.5	+ 0.3	+ 1.3	+ 0.7	+ 1.5	+ 0.7	+ 1.5

Table 7.6: Comparison of relative error reductions (in percentage) on telephone speech.

In Appendix C the WER and RER are given for every speaker of the test set.

ESST Experiments

The speech recognition experiments described in this section differ from those described in the last section inasmuch as the speech recognizer is trained on clean speech sampled at 16 kHz; testing was conducted on:

1. *clean speech*;
2. speech distorted by additive *white noise* and *robot noise*;
3. *meeting room* speech with significant reverberant distortion.

The results given in WER Table 7.7 and RER Table 7.8 show that the novel method of SMVSE outperforms its MVDR counterparts in all cased with two exceptions, the white noise at 6 dB and the meeting-room task. Applying SS the SMVDR outperforms all tested approaches.

	noise type *SNR (variance)*	*spectral* *subtraction*	**FFT-based** **MFCC**	**LP-based** **MFCC**	**MVDR-based** **MFCC**		**SMVDR-b.** **MFCC**
type				**autocorr.**	**autocorr.**	**Burg**	**autocorr.**
order				**20**	**120**	**120**	**120**
	clean speech	*without*	23.3	25.1	24.7	25.2	24.0
	white noise	*without*	46.9	46.7	45.1	45.9	46.5
	6 dB (6 dB)	*with*	42.4	43.4	43.1	43.4	41.2
	white noise	*without*	62.7	62.8	63.3	62.1	61.0
	4 dB (6 dB)	*with*	54.9	53.8	54.6	55.5	53.3
	white noise	*without*	72.5	73.8	74.5	73.9	72.8
	2 dB (6 dB)	*with*	64.6	62.0	63.5	63.6	61.0
	robot noise	*without*	40.2	42.4	41.2	40.3	42.5
	5 dB (7 dB)	*with*	38.7	45.1	43.0	42.5	37.8
	meeting-room	*without*	68.0	71.9	68.3	67.5	70.5
	10 dB (6 dB)	*with*	72.0	75.5	72.7	73.8	67.1
	delay&sum	*without*	47.8	56.6	49.3	48.9	48.2
	10 dB (6 dB)	*with*	55.5	67.1	59.8	58.1	54.5

Table 7.7: Comparison of word error rates (in percentage) in different acoustic conditions on continuous speech.

	noise type SNR (variance)	spectral subtraction	LP-based MFCC	MVDR-based MFCC		SMVDR-b. MFCC
type			autocorr.	autocorr.	Burg	autocorr.
order			20	120	120	120
	clean speech	without	- 7.7	- 6.0	- 8.2	- 3.0
	white noise	without	+ 0.4	+ 3.8	+ 2.1	+ 0.9
	6 dB (6 dB)	with	- 2.4	- 1.7	- 2.4	+ 2.8
	white noise	without	- 0.2	- 1.0	+ 1.0	+ 2.7
	4 dB (6 dB)	with	+ 2.0	+ 0.5	- 1.1	+ 2.9
	white noise	without	- 1.8	- 2.8	- 1.9	- 0.4
	2 dB (6 dB)	with	+ 4.0	+ 1.7	+ 1.5	+ 5.6
	robot noise	without	- 5.5	- 2.5	- 0.2	- 5.7
	5 dB (7 dB)	with	- 16.5	- 11.1	- 9.8	+ 2.3
	meeting-room	without	- 5.7	- 0.4	+ 0.7	- 3.7
	10 dB (6 dB)	with	- 4.9	- 1.0	- 2.5	+ 6.8
	delay&sum	without	- 18.4	- 3.1	- 2.3	- 0.8
	10 dB (6 dB)	with	- 20.9	- 7.7	- 4.7	+ 1.8

Table 7.8: Comparison of relative error reductions (in percentage) in different acoustic conditions on continuous speech.

In general, it can be seen that the SMVDR approach does relatively well for clean speech, only 0.7% worse in WER than the Fourier approach, and shows an overall performance which performs best in 7 out of 13 tested cases; see Figure 7.2. Considering only cases with additive noise, it even performs best in 5 out of 8 cases.

Computation Cost

Table 7.9 shows the computation costs of the different spectral estimation techniques compared to that of the FFT derived spectrum. While the autocorrelation and Burg approach remain in an acceptable range in comparison with the overall calculation time of the ASR system, the calculation cost of the modified covariance method is exorbitant. Moreover, the higher resolution provided by the modified covariance method is not expected to result in WER reductions. Hence, no experiments were conducted using this method

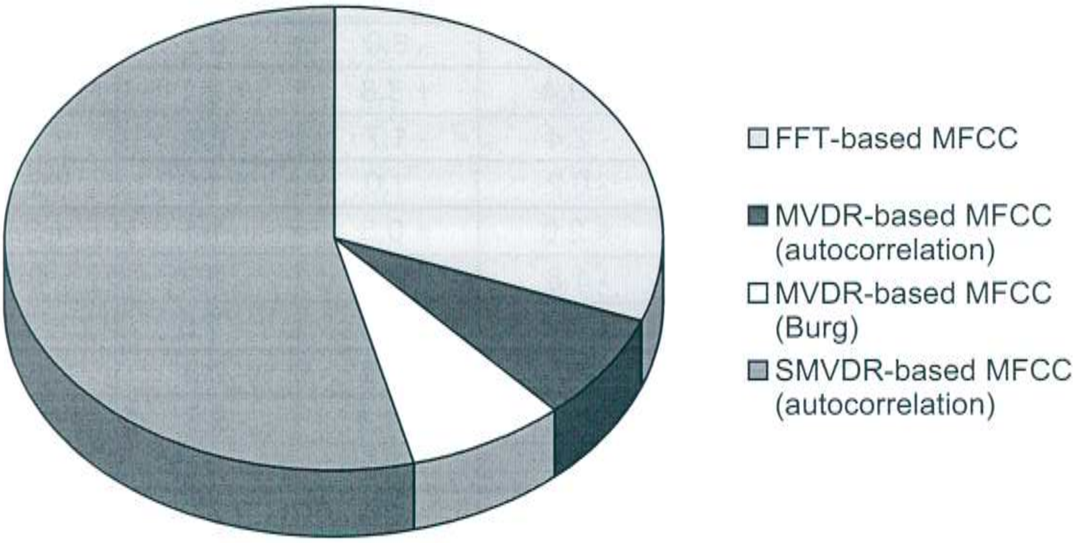


Figure 7.2: Overview in how many cases which approach performs best.

type	LP	scaled LP	MVDR		scaled MVDR	
order	20	20	60	120	60	120
autocorr.	1.81	2.98	3.03	5.44	4.20	6.61
Burg	2.10	3.27	4.79	8.50	5.96	9.67
mod. covar.	13.04	14.21	91.57	286.80	92.74	287.97
warp	3.03	4.20	5.91	10.77	6.98	11.94

Table 7.9: Time needed for computation in comparison to the computation time of the FFT

8 Conclusion and Future Work

8.1 Conclusion

In this thesis, novel feature extraction methods and their application to additive noise robustness in ASR have been investigated. We have seen that noise added to a signal distorts mainly the spectral valleys while the spectral peaks remain relatively unchanged. Therefore a SE instead of the Fourier spectrum was used to apply the feature extraction. To provide a good SE estimate, we have followed Dharanipragada and Rao [Dha01] in proposing the use of the MVDR instead of LP. One of the primary results of this thesis is a confirmation of their earlier finding that the spectral envelope estimate provided by MVDR is superior to that obtained with LP.

To further improve the robustness of the MVDR based approach to additive noise, a novel *scaling* method was introduced which adjusts the highest point of the envelope to the highest point of the Fourier spectrum. This could be shown to be, in average, less distorted by additive noise. Moreover, other well known techniques were adapted in this work: *Spectral subtraction* [Bol79] which could be shown to overcome the musical tones when applied on the envelope. *Pre-warping* [Mat98] which provides a better approximation of the aspects of the human auditory system than the envelope followed by a Mel-filterbank.

On *telephone speech*, SS did not lead to an improvement of accuracy, because the distortions seen in this speech are not, for the most part, additive in nature. Applying the proposed scaling of the MVDR derived envelope, however, lead to a slight improvement in accuracy. A further improvement could be obtained through the proposed pre-warping of the scaled MVDR envelope.

On *continuous speech* disturbed by additive noise, combining SS with the MVDR envelope and the proposed scaling could be shown to be superior to the Fourier transform, LP and

MVDR derive MFCC in most of the cases involving a mismatch between training and test conditions.

8.2 Future Work

8.2.1 Optimization of Parameters

Time constraints did not permit the several parameters pertaining to MVDR and SMVDR spectral estimation, such as the model order, the smoothing factor in SMVDR and the warping factor in WMVDR, to be tuned. Hence, further reductions in word error may come from optimizing these parameters empirically.

8.2.2 Mel Filterbank

The characteristic of a Mel filterbank results in different performances [Hyu99], which in the tested cases are optimized for the Fourier spectrum. For the MCDR-based MFCC, a particular filterbank refinement could lead to further improvements in the recognition performance. For example, it may prove beneficial to use a critical-band filter which is flat topped and non-symmetric as shown in Figure 8.1; such a filter is typical of the PLP approach [Mil02] to spectral estimation.

8.2.3 MVDR variations

As the MVDR approach is based on LPC, it opens up for a wide variety of suggested variations suggested and investigated for LP; e.g., an adaptation of PLP or Rasta-PLP to perceptual MVDR (PMVDR) or Rasta-PMVDR.

8.2.4 Speech Enhancement and Restoration of Wide Band Speech Signals from Narrow Band Speech Signals

In our investigations of the MVDR envelope and its variations, we saw that the MVDR approach successfully improves the robustness of an ASR system. Future work could concentrate on the enhancement of disturbed speech or on the restoration of wide band speech signals from narrow band speech signals to provide a more pleasant impression of the

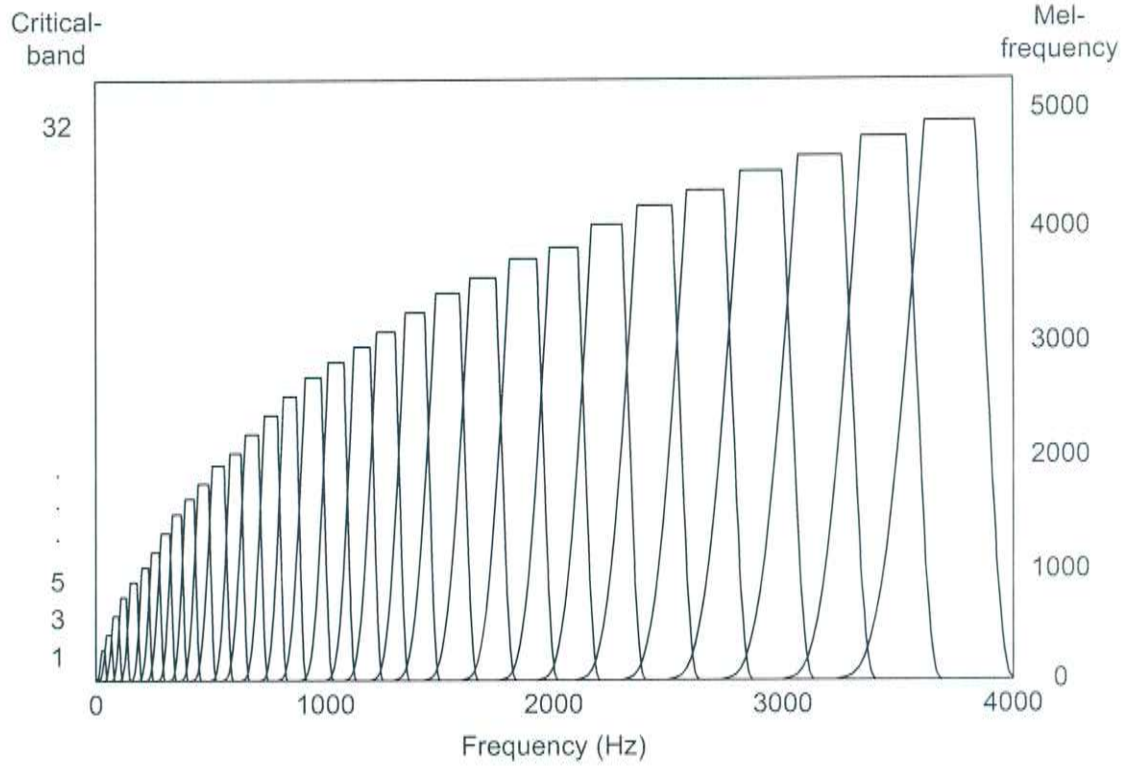


Figure 8.1: Critical band filters as applied in perceptual LP

speech to the human ear and an improvement of speech understanding for humans. In particular, it could be interesting in combination with SS, as it successfully prevents musical tones from occurring.

Appendix A

New JRTK Functions

A.1 SPECEST

The function *specest* estimates the SE of the PS using LP or MVDR through the autocorrelation method, the Burg method, the modified covariance method or the pre-warping approach.

A.1.1 Syntax

```
[FeatureSet] specest [Output] [Input] [Order] [Optional]
```

[FeatureSet]	Name of the featureset.
[Output]	Name of the output feature (type: FMatrix).
[Input]	Name of the input feature (type: FMatrix).
[Order]	Order of the prediction (typical values: LP 13..20, MVDR 60..120).
-type [Type]	Define the type of the estimate (LP or MVDR).
-lpmethod [Method]	Define the used method (autocorrelation, burg, modcovarianz or warp).
-correlate [Number]	Defines the number of used correlations. Only used by Burg and modified covariance. The number is calculated as windowsize (in ms) · samplingrate (in kHz).
-warp [Factor]	Warp factor of the warped spectral estimate

A.1.2 Example

This example calculates the MVDR spectrum using the Burg approach of order 120 all 10 ms using a window size of 20 ms.

```
set type MVDR
set method burg
set order 120
set wintype rect
set winsize 20
set shift 10
set filename e061ach1_034.16.shn

set random [expr{int(rand()*1000000)}]
set fs fs${random}

FeatureSet $fs
$fs readADC ADC1 $filename -bm shorten -v 0
$fs offset ADC ADC1 -mean 0 -alpha 0.02

set samplingrate [expr 1000*[$fs:ADC configure -samplingRate]]
set correlate [expr (samplingrate*winsize)]

$fs adc2spec ADC ${winsize}.ms -win $wintype -adc sADC \
    -shift ${shift}.ms
$fs specest SPEC sADC $order -type $type -lpmethod $method \
    -correlate $correlate
$fs log lSPEC SPEC 1.0 1.0

$fs show lSPEC
```

A.2 SPECADJ

The function *specadj* adjusts the hight of a feature given a second feature to be adjusted to. Useful to derive the SMVDR given the MVDR and FFT spectrum, see Example below.

A.2.1 Syntax

```
[FeatureSet] specadj [Output] [Input1] [Input2] [Optional]
```

[FeatureSet]	Name of the featureset.
[Output]	Name of the output feature (type: FMatrix).

[Input1] Name of the feature which should be adjusted (type: FMatrix).

[Input2] Name of the feature [Input1] should be adjusted to (type: FMatrix).

-smooth [Type] Smooth the [Input2] feature (0,1,2,3,4).

-show [Type] Show the multiplication factor for the scaling. (on,off)

A.2.2 Example

This example calculates the SMVDR spectrum by extending the last example with the following lines.

```
set blur 1

$fs spectrum FFT ADC $winsize -shift $shift

$fs specadj SPECadj SPEC FFT
$fs log lSPECadj SPECadj 1.0 1.0 -blur $blur

$fs show lSPECadj
```

Appendix B

Detailed WER & RER of Telephone Speech Experiments

speaker	FFT-b. MFCC	MVDR-b. MFCC		SMVDR-based MFCC							
type		auto.	Burg	autocorrelation				Burg			
order		60	60	60	80	100	120	60	80	100	120
1	17.9	19.4	19.4	16.9	16.9	19.4	18.1	20.4	18.3	19.0	21.9
2	27.9	28.7	26.2	28.6	28.4	27.8	27.4	27.9	28.6	27.8	30.0
3	22.4	25.0	24.8	23.0	24.1	23.4	24.1	21.8	23.7	24.9	26.0
4	31.1	32.8	33.1	30.8	29.7	29.1	29.4	32.0	30.8	29.9	31.7
5	30.3	31.2	29.1	29.8	28.9	29.2	28.0	29.1	29.2	31.6	29.1
6	38.4	35.8	36.1	37.4	35.5	38.2	36.1	37.6	34.7	36.1	36.8
7	41.5	38.9	38.9	41.1	41.3	41.5	42.0	41.3	42.6	42.6	43.5
8	32.3	30.4	33.2	32.9	32.1	33.2	34.2	36.4	35.3	35.9	35.6
9	43.8	43.8	42.6	41.2	45.0	44.9	44.0	44.1	44.9	42.6	42.6
10	21.4	22.8	25.3	24.2	22.4	24.0	23.2	23.2	25.3	25.1	25.7
11	36.0	37.9	37.3	36.9	35.5	36.9	36.0	37.1	35.8	36.0	37.1
12	45.3	45.1	43.8	41.7	41.7	42.1	44.7	43.2	43.4	46.5	45.9
13	36.1	35.4	37.0	36.6	37.0	37.8	37.2	37.5	36.2	38.3	37.5
14	38.3	41.4	40.6	39.1	38.9	37.7	37.7	40.3	40.6	41.7	40.6
15	66.8	69.7	68.5	65.6	63.7	68.5	66.3	70.9	70.4	63.2	66.8
16	65.6	65.3	62.3	68.9	65.8	64.4	59.7	63.2	64.4	65.1	64.6
average	36.2	36.8	36.4	36.2	35.9	36.5	35.9	36.8	36.8	36.9	37.5
mean	37.2	37.7	37.4	37.2	36.7	37.4	36.8	37.9	37.8	37.9	38.5

Table B.1: Comparison of word error rate (in percentage) on telephone speech.

speaker	MVDR-b. MFCC		SMVDR-based MFCC							
type	auto.	Burg	autocorrelation				Burg			
order	60	60	60	80	100	120	60	80	100	120
1	- 8.4	- 8.4	+ 5.6	+ 5.6	- 8.4	- 1.1	- 14.0	- 2.2	- 6.1	- 22.3
2	- 2.9	+ 6.1	- 2.5	- 1.8	+ 0.4	+ 1.8	+ 0.0	- 2.5	+ 0.4	- 7.5
3	- 11.6	- 10.7	- 2.7	- 7.6	- 4.5	- 7.6	+ 2.7	- 5.8	- 11.2	- 16.1
4	- 5.5	- 6.4	+ 1.0	+ 4.5	+ 6.4	+ 5.5	- 2.9	+ 1.0	+ 3.9	- 1.9
5	- 3.0	+ 4.0	+ 1.7	+ 4.6	+ 3.6	+ 7.6	+ 4.0	+ 3.6	- 4.3	+ 4.0
6	+ 6.8	+ 6.0	+ 2.6	+ 7.6	+ 0.5	+ 6.0	+ 2.1	+ 9.6	+ 6.0	+ 4.2
7	+ 6.3	+ 6.3	+ 1.0	+ 0.5	+ 0.0	- 1.2	+ 0.5	- 2.7	- 2.7	- 4.8
8	+ 5.9	- 2.8	- 1.9	+ 0.6	- 2.8	- 5.9	- 12.7	- 9.3	- 11.1	- 10.2
9	+ 0.0	+ 2.7	+ 5.9	- 2.7	- 2.5	- 0.5	- 0.7	- 2.5	+ 2.7	+ 2.7
10	- 6.5	- 18.2	- 13.1	- 4.7	- 12.1	- 8.4	- 8.4	- 18.2	- 17.3	- 20.1
11	- 5.3	- 3.6	- 2.5	+ 1.4	- 2.5	+ 0.0	- 3.1	+ 0.6	+ 0.0	- 3.1
12	+ 0.4	+ 3.3	+ 7.9	+ 7.9	+ 7.1	+ 1.3	+ 4.6	+ 4.2	- 2.6	- 1.3
13	+ 1.9	- 2.5	- 1.4	- 2.5	- 4.7	- 3.0	- 3.9	- 0.3	- 6.1	- 3.9
14	- 8.1	- 6.0	- 2.1	- 1.6	+ 1.6	+ 1.6	- 5.2	- 6.0	- 8.9	- 6.0
15	- 4.3	- 2.5	+ 1.8	+ 4.6	- 2.5	+ 0.7	- 6.1	- 5.4	+ 5.4	+ 0.0
16	+ 0.5	+ 5.0	- 5.0	- 0.3	+ 1.8	+ 9.0	+ 3.7	+ 1.8	+ 0.8	+ 1.5
average	- 1.7	- 0.6	+ 0.0	+ 0.8	- 0.8	+ 0.8	- 1.7	- 1.7	- 1.9	- 3.6
mean	- 1.3	- 0.5	+ 0.0	+ 1.3	- 0.5	+ 1.1	- 1.9	- 1.6	- 1.9	- 3.5

Table B.2: Comparison of relative error reductions (in percentage) on telephone speech.

Appendix C

Detailed WER & RER of Warp Experiments

speaker	FFT-b. MFCC	SMVDR-b. MFCC	SWMVDR-based MFCC					
type		auto.	autocorrelation					
order		80	40	50	60	70	80	90
1	22.3	20.0	20.4	20.8	20.2	20.2	21.9	19.6
2	29.0	28.3	27.9	28.8	28.2	27.4	28.8	29.5
3	26.8	26.8	25.9	27.5	26.1	25.9	25.7	25.6
4	32.8	32.6	31.7	31.7	31.7	30.5	30.2	30.8
5	31.4	33.7	30.1	32.3	31.0	31.9	30.5	32.4
6	40.0	39.2	41.1	41.3	41.3	41.1	42.9	44.2
7	42.8	41.1	42.4	42.4	43.5	42.4	43.3	42.4
8	35.6	38.6	36.1	37.0	37.0	34.2	36.7	34.2
9	46.9	44.9	45.8	46.1	47.3	46.0	46.1	47.6
10	23.8	24.4	23.0	20.5	23.0	22.4	22.8	21.6
11	39.8	37.7	39.2	38.9	38.5	40.6	37.9	39.4
12	47.6	47.6	49.5	43.0	47.6	48.8	46.5	46.7
13	40.7	40.7	38.9	39.6	40.8	40.7	42.1	40.2
14	41.7	47.1	49.4	45.4	44.9	44.0	45.7	44.3
15	70.4	69.0	70.6	71.8	70.9	69.0	70.9	69.7
16	68.4	65.6	66.5	64.2	63.0	65.8	63.5	62.5
average	39.1	38.7	38.7	38.4	38.7	38.5	38.7	38.5
mean	40.0	39.8	39.9	39.5	39.7	39.4	39.7	39.4

Table C.1: Comparison of word error rate (in percentage) of warped envelopes on telephone speech.

speaker	SMVDR-b. MFCC	SWMVDR-based MFCC					
type	auto.	autocorrelation					
order	80	40	50	60	70	80	90
1	+ 10.3	+ 8.5	+ 6.7	+ 9.4	+ 9.4	- 7.4	+ 5.8
2	+ 2.4	+ 3.8	+ 0.7	+ 2.8	+ 5.5	- 3.2	- 2.4
3	+ 0.0	+ 3.4	- 2.6	+ 2.6	+ 3.4	+ 0.8	+ 6.9
4	+ 0.6	+ 3.4	+ 3.4	+ 3.4	+ 7.0	+ 4.7	+ 2.8
5	- 7.3	+ 4.1	- 2.9	+ 1.3	- 1.6	- 1.3	- 0.3
6	+ 2.0	- 2.8	- 3.2	- 3.2	- 2.8	- 4.4	- 7.0
7	+ 4.0	+ 0.9	+ 0.9	- 1.6	+ 0.9	- 2.1	+ 0.0
8	- 8.4	- 1.4	- 3.9	- 3.9	+ 3.9	- 1.7	+ 7.6
9	+ 4.3	+ 2.3	+ 1.7	- 0.9	+ 1.9	- 0.7	- 3.3
10	- 2.5	+ 3.4	+ 13.9	+ 3.4	+ 5.9	+ 0.9	- 5.4
11	+ 5.3	+ 1.5	+ 2.3	+ 3.3	- 2.0	+ 3.3	- 1.3
12	+ 0.0	- 4.0	+ 9.7	+ 0.0	- 2.5	+ 6.1	- 8.6
13	+ 0.0	+ 4.4	+ 2.7	- 0.2	+ 0.0	- 8.2	- 1.5
14	- 12.9	- 18.5	- 8.9	- 7.7	- 5.5	+ 7.5	+ 2.4
15	+ 2.0	- 0.3	- 2.0	- 0.7	+ 2.0	- 0.4	+ 2.9
16	+ 4.1	+ 2.8	+ 6.1	+ 7.9	+ 3.8	+ 4.5	+ 2.6
average	+ 1.0	+ 1.0	+ 1.8	+ 1.0	+ 1.5	+ 0.0	- 0.3
mean	+ 0.5	+ 0.3	+ 1.3	+ 0.8	+ 1.5	+ 0.5	+ 0.3

Table C.2: Comparison of relative error reductions (in percentage) of warped envelopes on telephone speech.

Glossary

Notational Convention

$\mathbf{a}$	all vectors are column vectors and written in boldface
a_i	i^{th} element of $\mathbf{a}$
$a^{(i)}$	vector/prediction of order i
$\mathbf{a}^T$	transpose of vector
$\mathbf{A}$	all matrixes are capitalized and written in boldface
$A_{i,j}$	$[i, j]^{\text{th}}$ element of $\mathbf{A}$
$\mathbf{A}^T$	transpose of matrix
a^*	complex conjugate of a
$\hat{a}$	estimate of a
$\tilde{a}$	warped value of a
$\overline{a}$	smoothed value of a
$\sim$	"is distributed according to"
$/\cdot/$	denote the phoneme as a basic linguistic unit

Principal Symbols

a	linear prediction coefficient
$B(z)$,$b(n)$	backward prediction error
c	cepstral coefficient
d	spectral distortion
D	power difference
$D_k(z)$	Z-transform of a k^{th} sub-filter
f	frequency

E_p	mean-square prediction error
$E(z)$,$e(n)$	(forward) prediction error
$\mathbf{h}$	tab-weight vector
$H(z)$	transfer function of discrete-time linear filter
k	partial correlation coefficient, reflection coefficient
M	final order
n	element n of a discrete time series
N	data length, noise
$\mathbf{R}$	correlation matrix
$\mathbf{R}_{xx}$	autocorrelation matrix
s	fixed frequency vector
$S(\omega)$	power spectral density
$S_{LP}(\omega)$	linear prediction spectrum
$S_{MV}(\omega)$	minimum variance distortionless response spectrum
$w(n)$	window function
ϵ	modelling error
λ	warping parameter
μ	correlated prediction coefficient
ϕ	short-time autocorrelation function
ω	frequency
ω_{foi}	frequency of interest

Abbreviations

AR	AutoregRessive
ARMA	AutoregRessive Moving Average
ASR	Automatic Speech Recognition
dB	deciBel
DCT	Discrete Cosine Transformation
ESST	English Spontaneous Scheduling Task
EMNS	Ephraim and Malah Noise Suppressor

FFT	Fast Fourier Transform
FOI	Frequency Of Interest
HMM	Hidden Markov Model
JRTK	Janus Research ToolKit
LDA	Linear Discriminant Analysis
LM	Lattice Method
LP	Linear Prediction
LPC	Linear Prediction Coefficient
LPCC	Linear Prediction Cepstral Coefficient
LPS	Logarithmic Power Spectrum
LSD	Logarithmic Spectral Distortion
LSF	Logarithmic Spectral Feature
MA	Moving Average
MFCC	Mel-Frequency Cepstral Coefficients
ML	Maximum Likelihood
MLSF	Mel-filtered Log Spectral Feature
MMSE	Minimum Mean Squared Error
MV	Minimum Variance
MVDR	Minimum Variance Distortionless Response
MVDRCC	Minimum Variance Distortionless Response Cepstral Coefficients
MVSE	Minimum Variance Spectral Estimation
PARCOR	Partial Correlation
PLP	Perceptual Linear Predictive
PMVDR	Perceptual Minimum Variance Distortionless Response
PS	Power Spectrum
PSD	Power Spectral Density
RER	Relative Error Reduction
RSR	Robust Speech Recognition
RTF	Real Time Factor
SE	Spectral Envelope
SFFT	Smoothed Fast Fourier Transform
SMVDR	Scaled Minimum Variance Distortionless Response

SMVSE	Scaled Minimum Variance Spectra Estimation
SNR	Signal-to-Noise Ratio
SP	Spectral Distortion
SS	Spectral Subtraction
VTLN	Vocal Tract Length Normalization
WA	Word Accuracy
WER	Word Error Rate
WLP	Warped Linear Prediction
WMVDR	Warped Minimum Variance Distortionless Response

Bibliography

[Ars95] ARSLAN, L.; MCCREE, A. and VISWANATHAN, V.: *New methods for adaptive noise suppression*. ICASSP, vol. 1:pp. 812–815, 1995.

[Bar97] BARKER, J. and COOKE, M.P.: *Modelling the recognition of spectrally reduced speech*. Eurospeech, pp. 2127–2130, 1997.

[Bat98] BATRI, N.: *Robust spectral parameter coding in speech processing*. Master Thesis, McGill University, Montreal, Canada, May 1998.

[Bol79] BOLL, S.F.: *Suppression of acoustic noise in speech using spectral subtraction*. ASSP, vol. 27:pp. 113–120, Apr. 1979.

[Bro95] BRONSTEIN, N.; SEMENDJAJEW, K.A.; MUSIOL, G. and MÜHLIG, H.: *Taschenbuch der Mathematik*. Verlag Harri Deutsch, 1995.

[Bur72] BURG, J.P.: *The relationship between maximum entropy and maximum likelihood spectra*. Geophysics, vol. 37:pp. 375–376, Apr. 1972.

[Cap69] CAPON, J.: *High-resolution frequency-wavenumber spectrum analysis*. Proc. IEEE, vol. 57:pp. 1408–1418, August 1969.

[Cap94] CAPPÉ, O.: *Elimination of the musical noise phonemenon with the Ephraim and Malah noise suppressor*. SAP, vol.2(no.2), April 1994.

[Coo97] COOKE, M.; MORRIS, A. and GREEN, P.: *Missing data techniques for robust speech recognition*. ICASSP, pp. 863–866, 1997.

[Del00] DELLER, J.R.; HANSEN JR., J.H.L. and PROAKIS, J.G.: *Discrete-time processing of speech signals*. IEEE Press, 2000.

[Del93] DELLER JR., J.R.; PROAKIS, J.G. and HANSEN, J.H.L.: *Discrete-time processing of speech signal*. Macmillan, 1993.

[Dha01] DHARANIPRAGADA, S. and RAO, B.D.: *MVDR based feature extraction for robust speech recognition*. ICASSP, vol.1:pp. 309–312, 2001.

[Dim89] DIMOLITSAS, S.: *Objective speech distortion measures and their relevance to speech quality assessments.* IEEE Proc. I Communications, Speech and Vision, vol. 136:pp. 317–324, Oct. 1989.

[Dim95] DIMOLITSAS, S.; CORCORAN, F.L. and RAVISHANKAR, C.: *Dependence of opinion scores on listening sets used in degradation category rating assessments.* ASSP, vol. 3:pp. 421–424, Sept. 1995.

[Dob95] DOBLINGER, G.: *Computationally efficient speech enhancement by spectral minima tracking in subbands.* Eurospeech, vol. 2:pp. 1513–1516, 1995.

[Don00] MCDONOUGH, J.W.: *Speaker compensation with all-pass transforms.* Doctor Thesis, Johns Hopkins University, Baltimore, Maryland, USA, 2000.

[Edl00] EDLER, B. and SCHULLER, G.: *Audio coding using a psychoacustic pre- and postfilter.* ICASSP, vol. 2:pp. 881–884, 2000.

[Eph84] EPHRAIM, Y. and MALAH, D.: *Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator.* ASSP, vol. 32(no. 6):pp. 1109–1121, Dec. 1984.

[Gel02] GELBART, D. and MORGAN, N.: *Double the trouble: Handling noise and reverberation in far-field automatic speech recognition.* ICSLP, Sep. 2002.

[Goh72] GOHBERG, I.C. and SEMENCUL, A.A.: *On the inversion of finite Toeplitz matrices and their continuous analogs.* Mat. Issled, vol. 2:pp. 201–233, 1972.

[Gu01] GU, L. and ROSE, K.: *Perceptual harmonic cepstral coefficients for speech recognition in noisy environment.* ICSSP, vol. 1:pp. 125–128, 2001.

[Hay02] HAYKIN, S.: *Adaptive filter theory—4th ed.* Prentice Hall, 2002.

[Hay91] HAYKIN, S.: *Adaptive filter theory—3th ed.* Prentice Hall, 1991.

[Her90] HERMANSKY, H.: *Perceptual linear predictive (PLP) analysis of speech.* ASA, 87:pp. 1738–1752, 1990.

[Hol01] HOLMES, J. and HOLMES, W.: *Speech synthesis and recognition.* Taylor and Francis, 2001.

[Hos79] HOSOYA, Y.: *Efficient estimate of a model with an autoregressive signal with white noise.* Technical Report Tech. Rep. 37, Department of Statistics, Standford University, Mar. 1979.

[Hua01] HUANG, X.; ACERO, A. and HON, H.W.: *Spoken language processing*. Prentice Hall, 2001.

[Hyu99] HYUN, D. and LEE, C.: *Optimization of Mel-cepstrum for speech recognition*. IEEE International Conference on Systems, Man, and Cybernetics, vol. 1:pp. 500–503, 1999.

[Ita68] ITAKURA, F. and SAITO, S.: *Analysis synthesis telephony based on the maximum likelihood method*. Proc. 6th Int. Cong. Acoust., C-5-5, 1968.

[Ita71] ITAKURA, F. and SAITO, S.: *Digital filtering techniques for speech analysis and synthesis*. Proc. 7th Int. Cong. Acoust., 25-C-1:pp. 261–264, 1971.

[Jac96] JACKSON, B. and LELAND, ?: *Digital filters and signal processing: with Matlab exercises*. Kluwer Academic Publishers, 1996.

[Jar91] EL-JAROUDI, A. and MAKHOUL, J.: *Discrete all-pole modeling*. IEEE Trans. Speech Processing, vol. 39:pp. 411–423, Feb. 1991.

[Jel99] JELINEK, M. and ADOUL, J.P.: *Frequency-domain spectral envelope estimation for low rate coding of speech*. ICASSP, pp. 253–256, 1999.

[Kab00] KABAL, P. and KLEIJN, B.: *All-pole modelling of mixed excitation signals*. ICASSP, vol. 1:pp. 97–100, 2000.

[Kai78] KAILATH, T.; VIEIRA, A. and MORF, M.: *Inverses of Toeplitz operators, innovations, and orthonormal polynomials*. SIAM Rev., vol. 20:pp. 106–119, Jan. 1978.

[Kam98] KAMMEYER, K.D. and KROSCHEL K.: *Digitale Signalverarbeitung: Filterung und Spektralanalyse mit Matlab-Übungen*. B.G. Teubner, 1998.

[Kar01] KARJALAINEN, M.: *Auditory interpretation and application of warped linear prediction*. Proceedings of Consistent & Reliable Acoustic Cues for Sound Analysis, 2001.

[Kay88] KAY, S.M.: *Modern spectral estimation: Theory and application*. Englewood Cliffs, Prentice Hall, 1988.

[Kon94] KONDOZ, A.M.: *Digital speech Coding for low bit rate communications systems*. John Wiley and Sons, 1994.

[Kot01] KOTNIK, B.; KACIC, Z. and HORVAT, B.: *A multiconditional robust front-end feature extraction with a noise reduction procedure based on improved spectral subtraction*. Eurospeech, 2001.

[Lag84] LAGUNAS, M.A. and GASULL, A.: *An improved maximum likelihood method for power spectral density estimation*. ASSP, vol. 32:pp. 170–173, Feb. 1984.

[Lee96] LEE, C.H.; SOONG, F.K. and PALIWAL, K.K.: *Automatic speech and speaker recognition*. Kluwer Academic Publishers, 1996.

[Li98] LI, H.; STOICA, P. and LI, J.: *Capon estimation of covariance sequences*. Circuits, Systems, and Signal Processing, vol. 17(no. 1):pp. 29–49, Jan. 1998.

[Lim78] LIM, J.S.: *All pole modeling of degraded speech*. ASSP, vol. 26:pp. 197–209, June 1978.

[Lim96] LIM, I.T. and LEE, B.G.: *Lossy pole-zero modeling for speech signals*. SAP, vol. 4:pp. 81–88, Mar. 1996.

[Lin97] LINHARD, K. and KLEMM, H.: *Noise reduction with spectral subtraction and median filtering for suppression of musical tones. In Robust speech recognition using unknown communication channels*. ESCA-NATO Tutorial and Research Workshop, pp. 159–162, April 1997.

[Lip97] LIPPMANN, R.P. and CARLSON, B.A.: *Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering, and noise*. Eurospeech, (No. KN-37), September 1997.

[Lom11] LOMBARD, E.: *Le signe de l'élévation de la voix*. Ann. Maladies Oreille, Larynx, Nez, Pharynx, vol. 37:pp. 101–119, 1911.

[Mak73] MAKHOUL, J.: *Spectral analysis of speech by linear prediction*. ASSP, vol. 21(no. 3):pp.140–148, 1973.

[Mar01] MARTIN, R.: *Noise power spectral density estimation based on optimal smoothing and minimum statistics*. SAP, vol. 9(no. 5):pp. 504–512, July 2001.

[Mar94] MARTIN, R.: *Spectral subtraction based on minimum statistics*. European Signal Processing Conference, pp. 1182–1185, 1994.

[Mat01] MATSUMOTO, H. and MOROTO, M.: *Evaluation of Mel-LPC cepstrum in a large vocabulary continuous speech recognition*. ICASSP, vol. 1:pp. 117–120, 2001.

[Mat98] MATSUMOTO, M.; NAKATOH, Y. and FURUHATA, Y.: *An efficient Mel-LPC analysis method for speech recognition*. ICSLP, pp. 1051–1054, 1998.

[Mey97] MEYER, J.; SIMMER, K.U. and KAMMEYER, K.D.: *Comparison of one and two-channel noise-estimation techniques*. Proc. Int. Workshop Acoustic Echo Control Noise Reduction, pp. 17–20, 1997.

[Mil02] MILNER, B.: *A comparison of front-end configurations for robust speech recognition*. IEEE, 2002.

[Mol01] MOLAU, S.; PITZ, M.; SCHLÜTER, R. and NEY H.: *Computing Mel-frequency cepstral coefficients of the power spectrum*. ICASSP, 2001.

[Mor88] MORII, S.: *Spectral subtraction in the sphinx system (unpublished)*. 1988.

[Mur00] MURTHI, M.N. and RAO, B.D.: *All-pole modeling of speech based on the minimum variance distortionless response spectrum*. ICASSP, vol. 8(no. 3):pp. 221–239, May 2000.

[Mur97] MURTHI, M.N. and RAO, B.D.: *All-pole model parameter estimation for voiced speech*. IEEE Workshop Speech Coding Telecommunications Proc., Pacono Manor, PA, 1997.

[Mus85] MUSICUS, B.R.: *Fast MLM power spectrum estimation from uniformly spaced correlations*. ASSP, vol. 33:pp. 1333–1335, 1985.

[Opp89] OPPENHEIM, A.V. and SCHAFER, R.W.: *Discrete-time signal processing*. Prentice-Hall Inc., 1989.

[Pal93] PALIWAL, K.K. and ATAL, B.S.: *Efficient vector quantization of LPC parameters at 24 bits/frame*. IEEE Trans. Speech and Audio Proc., vol. 1:pp. 3–14, January 1993.

[Pri81] PRIESTLEY, M.B.: *Spectral analysis and time series, vol.2 of probability and mathematical statistics*. Academic Press Inc., London, 1981.

[Qua88] QUACKENBUSH, S.R.; BARNWELL, T.P. and CLEMENTS, M.A.: *Objective measures of speech quality*. Englewood Cliffs, Prentice Hall, 1988.

[Rab78] RABINER, L.R. and SCHAFER, R.W.: *Digital processing of speech signals*. Prentice-Hall Inc., 1978.

[Rab93] RABINER, L. R. and JUANG, B. H.: *Fundamentals of speech recognition*. Englewood Cliffs, New Jersey: Prentice-Hall, 1993.

[Sai85] SAITO, S. and NAKATA, K.: *Fundamentals of speech signal processing*. Academic Press, 1985.

[Sch75] SCHWARTZ, R. and MAKHOUL, J.: *Where the phonemes are: Dealing with ambiguity in acoustic-phonetic recognition.* ASSP, vol. 23:pp. 50–53, 1975.

[Sch98] SCHWARZ, D.: *Spectral envelopes in sound analysis and synthesis.* Diploma Thesis Nr. 1622, Universität Stuttgart, Fakultät Informatik, Germany, 1998.

[Sha00] O'SHAUGHNESSY, D.: *Speech communications: Human and machine.* IEEE Press, 2000.

[Sha87] O'SHAUGHNESSY, D.: *Speech communication: Human and machine.* Addison-Wesley Publishing Company, 1987.

[Sri79] SRINATH, M. and RAJASEKARAN, P.: *An introduction to statistical signal processing with applications.* John Wiley and Sons, 1979.

[Str80] STRUBE, H.W.: *Linear prediction on a warped frequency scale.* ASA, vol. 68(no. 8):pp. 1071–1076, 1980.

[Str90] STROBACH, P.: *Linear prediction theory: A mathematical basis for adaptive systems.* Springer-Verlag, 1990.

[Tie80] TIERNEY, J.: *A study of LPC analysis of speech in additive noise.* ASSP, vol. 28(no. 4), 1980.

[Tok95] TOKUDA, K.; KOBAYASHI, T. and IMAI, S.: *Adaptive cepstral analysis of speech.* SAP, vol. 3(no. 6):pp. 481–489, 1995.

[Vas92] VASEGHI, S. and FRAYLING-CORK, R.: *Restoration of old gramophone recordings.* J. Audio Eng., vol. 40(no. 10):pp. 791–801, 1992.

[Wet00] DE WET, F.; CRANEN, B.; DE VETH, J. and BOVES, L.: *Comparing acoustic features for robust ASR in fixed and cellular network applications.* ICASSP, pp. 1415–1418, 2000.

[Wet01] DE WET, F.; CRANEN, B.; DE VETH, J. and BOVES, L.: *A comparison of LPC and FFT-based acoustic features for noise robust ASR.* Eurospeech, 2001.

[Wie66] WIENER, N.: *Extrapolation interpolation and smoothing of stationary time series.* MIT Press, 1966.

[Wil02] WILLIAMS, M.: *Sony shows off humanoid robot - the SDR-4X moves gracefully, and can understand complex speech and facial expressions.* http://cssvc.pcworld.compuserve.com/computing/cis/article/0,aid,89763,00.asp, 2002.

Conferences

Eurospeech	European Conference on Speech-Communication and Technology
ICASSP	IEEE International Conference on Acoustic, Speech, and Signal Processing
ICSLP	International Conference on Speech and Language Processing

Journals

ASSP	IEEE Transactions on Acoustics, Speech and Signal Processing
ASA	Journal Acoustic Society of America
SAP	IEEE Transactions on Speech and Audio Processing

Index