HELSINKI UNIVERSITY OF TECHNOLOGY
Faculty of Information and Natural Sciences
Department of Information and Computer Science

Heikki Kallasjoki

Methods for Spectral Envelope Estimation in Noise Robust Speech Recognition

Master's Thesis
Espoo, August 26, 2009

Supervisor:	Professor Erkki Oja, Helsinki University of Technology
Instructor:	Kalle Palomäki, D.Sc. (Tech.), Helsinki University of Technology

HELSINKI UNIVERSITY OF TECHNOLOGY
Faculty of Information and Natural Sciences
Degree Programme of Computer Science and Engineering

ABSTRACT OF
MASTER'S THESIS

Author:	Heikki Kallasjoki		
Title of thesis:	Methods for Spectral Envelope Estimation in Noise Robust Speech Recognition		
Date:	August 26, 2009	**Pages:**	10 + 62
Professorship:	Computer and Information Science	**Code:**	T-61
Supervisor:	Professor Erkki Oja		
Instructor:	Kalle Palomäki, D.Sc. (Tech.)		

In real world applications speech recognition systems are challenged by a variety of noisy environments, where prior knowledge of the type of noise may not be available. The short-time spectral envelope of a speech signal embodies the information relevant to the linguistic message communicated with the speech in a form that is more robust against noise than the underlying spectrum. Spectral envelope estimates can therefore be exploited to produce feature representations for speech signals that are less affected by environmental noise.

In this work, different spectral envelope estimation methods are evaluated in the feature extraction stage of a large vocabulary continuous speech recognition system. Recognition error rates for speech recorded in real-world noisy environments are compared between feature representations based on conventional, weighted and a recently proposed stabilized weighted linear predictive spectral envelope estimate, as well as FFT and other baseline methods. Methods for automatically adapting the parameters of the stabilized weighted linear prediction method to the analyzed audio data are also investigated. Significantly better recognition results are obtained using the systems based on conventional and weighted linear predictive spectral envelope estimates compared to the baseline FFT system, when recognizing noisy speech using acoustic models trained with clean speech.

Keywords:	automatic speech recognition, feature extraction, spectral envelope estimation, weighted linear prediction
Language:	English

TEKNILLINEN KORKEAKOULU
Informaatio- ja luonnontieteiden tiedekunta
Tietotekniikan koulutusohjelma

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä:	Heikki Kallasjoki		
Työn nimi:	Menetelmiä spektrin verhokäyrän mallintamiseen kohinasietoista puheentunnistusta varten		
Päiväys:	26. elokuuta 2009	**Sivumäärä:**	10 + 62
Professuuri:	Informaatiotekniikka	**Koodi:**	T-61
Työn valvoja:	Professori Erkki Oja		
Työn ohjaaja:	TkT Kalle Palomäki		

Puheentunnistusjärjestelmien käytännön sovellusten on toimittava haasteellisissa ympäristöissä, joissa tietoa mahdollisten häiriöäänien laadusta ei välttämättä ole saatavilla ennakkoon. Puhesignaalin lyhytaikaisen spektrin verhokäyrä sisältää puheen välittämään viestiin liittyvän tiedon muodossa, joka sietää kohinaa paremmin kuin sen perustana oleva spektri. Malleja verhokäyrästä voidaankin siten käyttää tuottamaan puhesignaaleille piirre-esitystapoja, joihin ympäristöstä peräisin olevat häiriöäänet vaikuttavat vähemmän.

Tässä työssä eri menetelmiä verhokäyrän mallintamiseen tarkastellaan osana laajan sanavaraston jatkuvan puheen tunnistusjärjestelmän piirreirroitusta. Tunnistusvirheiden määrää todellisissa kohinaisissa ympäristöissä nauhoitetulle puheelle vertaillaan eri piirre-esitystapojen välillä. Vertailtavat piirteet perustuvat perinteiseen, painotettuun ja hiljattain esiteltyyn stabiloituun painotettuun lineaariprediktioon. Myös menetelmiä stabiloidun painotetun lineaariprediktion parametrien automaattiseen mukautukseen analysoitavalle äänisignaalille sopivaksi tutkitaan. Merkittävästi parempia tunnistustuloksia saavutetaan perinteiseen ja painotettuun lineaariprediktioon perustuvia verhokäyrän malleja käyttävillä piirteillä, kun tunnistettavana on kohinaista puhetta ja käytetään puhtaalla puheella opetettuja akustisia malleja.

Avainsanat:	automaattinen puheentunnistus, piirreirroitus, spektrin verhokäyrän mallinnus, painotettu lineaariprediktio
Kieli:	englanti

Acknowledgements

The work reported in this thesis was done in the Department of Information and Computer Science at the Helsinki University of Technology during the years 2007, 2008 and 2009, and supported by the Academy of Finland project *Auditory approaches to automatic speech recognition*.

I thank professor Erkki Oja for supervising the thesis. I thank my instructor Kalle Palomäki for the opportunity to work in the speech group, and valuable feedback on the contents of this thesis. I am also grateful to everyone in the speech recognition group for both the prior work upon which this thesis is built, as well as their help in everything related to said work. I would also like to extend thanks to our collaborators at the Department of Signal Processing and Acoustics, especially Paavo Alku, Carlo Magi and Jouni Pohjalainen, for providing the original inspiration for this thesis.

Finally, I thank my family and friends for their support, with a special note of thanks to my wife for reading drafts of this thesis and tolerating some very late Sunday evenings spent working on it.

Espoo, August 26, 2009

Heikki Kallasjoki

Contents

List of Tables

List of Figures

Symbols and abbreviations

$\mathbf{a}$	Linear prediction coefficients
M	Short-time energy window width
$\mathbf{R}$	Autocorrelation matrix
$s(n)$	Speech audio signal
$\mathbf{s}(\tau)$	Observed speech features at time τ
$\mathbf{S}$	Sequence of observed speech features
$p(W\|\mathbf{S})$	Probability for a word sequence W given observations $\mathbf{S}$
$p(W)$	Prior probability for a word sequence W
w_n	Weight function values

ASR	Automatic speech recognition
CMS	Cepstral mean subtraction
DCT	Discrete cosine transform
FFT	Fast Fourier transform
FIR	Finite impulse response
GMM	Gaussian mixture model
HMM	Hidden Markov model
LER	Letter error rate
LP	Linear prediction
LVCSR	Large vocabulary continuous speech recognition
MFCC	Mel-frequency cepstral coefficients
MLLR	Maximum likelihood linear regression
MLLT	Maximum likelihood linear transform
MVDR	Minimum variance distortionless response
PLP	Perceptual linear prediction
STE	Short-time energy
SNR	Signal-to-noise ratio
SWLP	Stabilised weighted linear prediction
WER	Word error rate
WLP	Weighted linear prediction

Chapter 1

Introduction

Speech is for the most a natural means of communication, and certainly it would be advantageous if we could use spoken language as the medium also when interfacing with computer systems. Automatic speech recognition (ASR) systems have therefore been developed to enable computers to algorithmically convert a digital audio signal to text. So far these systems have mostly been confined to uses where the recognition vocabulary is limited, such as command-based user interfaces for different devices, but research continues on the challenging task of recognizing fluent speech with unlimited vocabulary.

As speech has evolved for communication between humans, we have a natural aptitude for it, with the capability of adapting to different speakers and auditory environments. Automatic speech recognition systems, however, have difficulties dealing with changes to the speaker or the environment. This poses a problem for real-world applications such as mobile computing, where a speech recognition system is used in a variety of noisy environments, with no prior knowledge of potential noise types available.

Computer processing of speech utilizes digitized audio signals. The digitized audio waveform itself, however, is not suitable for speech recognition use directly, as the information about the spoken words it contains is not in an easily extractable form. In the feature extraction stage the audio signal is therefore transformed into a sequence of feature vectors designed to capture the salient information related to the text content in a usable way.

In this work, the focus is on a specific class of methods for enhancing noise robustness in speech recognition: the use of spectral envelope estimates in the feature extraction stage, motivated by the fact that the short-time spectral envelope captures the information relevant to the text content of the speech in a more noise-robust way than the underlying spectrum. Specifically, we have chosen to evaluate the speech recognition performance of feature extraction systems based on conventional [29], weighted [26] and recently proposed stabilized weighted [28] linear prediction signal modeling methods. The perceptual linear prediction [15] method, as well as various methods based on the minimum variance distortionless response [10, 31, 52] are also considered. The popular mel-frequency cepstral coefficient [8] feature extraction scheme is used as a baseline reference for all the methods.

The primary goal of the research on which this thesis is based was to evaluate the suitability of the stabilized weighted linear prediction (SWLP) method for speech recognition tasks, in collaboration with the Department of Signal Processing and Acoustics of Helsinki University of Technology where the method was originally developed. To this end, a set of speech recognition experiments were designed and performed. In this thesis, the results of these experiments as well as our further research on the adaptive selection of optimal parameters for the SWLP algorithm are reported.

The speech recognition experiments are carried out using the large vocabulary continuous speech recognition (LVCSR) system developed in the Adaptive Informatics Research Centre at Helsinki University of Technology. The recognizer is a state of the art hidden Markov model based system, using Gaussian mixture models and Gamma distribution duration modeling for the acoustic models, a variable-length sub-word n-gram language model and a single-pass time-synchronous Viterbi decoder [7, 17, 18, 36, 37, 43]. The SPEECON Finnish language corpus [20] was used as the source for real-world noisy speech for training and recognition.

The structure of the thesis is as follows. An introduction to the topics of automatic speech recognition and spectral envelope estimation is given in Chapter 2. The feature extraction process and the evaluated spectral envelope estimation methods are presented in more detail in Chapter 3. In Chapter 4 a series of speech recognition experiments are performed using the methods, and Chapter 5 discusses the results of these experiments. Chapter 6 concludes the thesis by considering the stated research questions in light of the body of this work.

Chapter 2

Automatic speech recognition

Large vocabulary continuous speech recognition (LVCSR) is one of the most challenging and fundamental tasks in the field of automatic speech recognition. The large vocabulary size makes it difficult to distinguish between individual words, many of which may have very similar phonetic structure. In addition, it is more difficult to construct a sufficiently large set of training data in order to have representative samples available for each word. In continuous, fluent speech there are not always pauses to indicate word borders, and the amount of possible interpretations for the audio data grows rapidly, making it harder to find the correct one. Coarticulation between words can also cause significant changes in the acoustic representation of a single word. In this section, the general structure of an LVCSR system is described. For the most part, the discussion is based on [9, 14, 19, 39].

The fundamental task of a speech recognition system is to find the word sequences that are the most likely to have produced the observed acoustic information. Figure 2.1 illustrates the speech recognition process. The initial observed audio signal waveform $x(n)$ is not suited for speech recognition as-is, as the acoustic information related to the spoken words is represented in a complex way, and the signal also contains information irrelevant from a speech transcription point of view. The audio signal is therefore transformed to a sequence of feature vectors $\mathbf{S} = \{\mathbf{s}(\tau)\}$, where $\mathbf{s}(\tau)$ are the features observed at a discrete time τ, using a suitable time step so that the spectral properties of speech are approximately stationary for each frame.

The task of the decoder is to find the most likely word sequence $\hat{W}$ based on the posterior probability $P(W|\mathbf{S})$, the conditional probability of the word

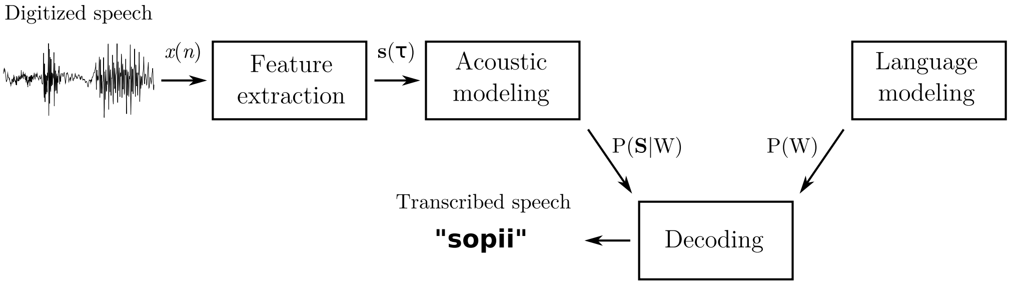

Figure 2.1: *The automatic speech recognition process*

sequence W given the observed speech features $\mathbf{S}$. Acoustic models are used to estimate the probability $P(\mathbf{S}|W)$ of observing the particular sequence of feature vectors, given the word sequence W. The prior probability $P(W)$ for the word sequence is given by the language model. The most likely word sequence $\hat{W}$ can then be obtained via Bayes' theorem from the model probabilities as [9]

$$\hat{W} = \arg\max_{W} P(W|\mathbf{S}) = \arg\max_{W} P(\mathbf{S}|W)\, P(W). \tag{2.1}$$

Each step is described in more detail in sections 2.1 to 2.4.

2.1 Feature extraction

The task of the feature extraction stage is to capture from the acoustic signal the information relevant for discriminating between the spoken words or phonemes, while discarding any irrelevant information, such as speaker-dependent variation and the prosodic content of speech. As speech has evolved as a communication mechanism between humans, studies of the traits of the human speech production and auditory system have been used as guidance when developing feature extraction methods. Successful feature extraction methods have usually included these psychoacoustical considerations (e.g. [8, 10, 15], among others).

Due to physical limitations of the human speech production system, the rate of change of the phonetic features of speech is limited. The frequency components of the speech signal are therefore nearly stationary over a short period of time, and the short-time frequency spectrum is commonly used as the basis for the feature vector representation for speech.

Mel-frequency cepstral coefficients (MFCCs) [8] are a popular choice for a feature representation in ASR systems. To approximate the spectral resolution of the human hearing, the nonlinear mel frequency scale is used. The speaker-dependent variation is also partially suppressed by the MFCC features. A thorough description of the MFCC feature extraction system used in this work is given in Section 3.2.

The upper envelope of the short-time spectrum contains the information relevant to the spoken phonemes, and can therefore be used to construct more noise-robust speech features. The topic of spectral envelope estimation is discussed in Section 3.1, and the estimation methods used in this work in sections 3.3 to 3.7.

2.2 Acoustic modeling

Phonemes are the basic units used to construct words in spoken language, analogously to the role of letters in written text. To estimate the probability of observing a sequence of speech features, given a sequence of words, we have to establish a correspondence between the phonemes of the word sequence and the individual feature vectors. However, the temporal duration region of a single phoneme will cover a chain of feature vectors, and the phonemes themselves have an internal time structure. Furthermore, the pronunciation of a phoneme depends on its context of adjacent phonemes, causing a coarticulation effect. The acoustic models need to account for all these complications. Hidden Markov models (HMM) are the predominant method for acoustic modeling in modern ASR systems [13].

A discrete-time first order Markov chain is a statistical system, which at any time is in one of a set of N states, and can change its state at discrete time steps. Denoting the state of the system at time t as q_t, the probability $P(q_t = S_i)$ for the system to be in a particular state S_i at a given time depends only on the previous state,

$$P(q_t = S_i | q_{t-1} = S_j, q_{t-2} = S_k, \dots) = P(q_t = S_i | q_{t-1} = S_j). \tag{2.2}$$

The system is therefore perfectly described by the static *state transition probabilities*, denoted a_{ij}, between the states.

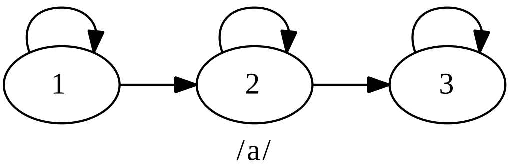

Figure 2.2: *Allowed transitions in a three-state HMM for the phoneme /a/*

In the hidden Markov model case, the model states themselves are not observed. Instead, the system creates a sequence of observations $\mathbf{O} = \{\mathbf{o}(\tau)\}$, based on per-state probability distributions called *observation probabilities* for the observed values, $b_i(\mathbf{o})$. The same particular observation can be generated by different model states, each of which has a separate observation probability for it. When used in acoustic modeling for speech recognition, the observations generated by an HMM correspond to the feature vectors of recorded speech, while the state sequence corresponds to the sequence of phonemes or other acoustic units, with single phonemes often modeled with multiple states. The task is then to find the underlying state trajectory responsible for producing the observed features.

In general, the states and the allowed transitions between states are fixed beforehand for the HMM, and in model training simply the state transition and observation probabilities are estimated. As a single HMM for a phoneme cannot capture the context-dependent variation in pronunciation, triphones are often chosen as the basic unit for modeling in large vocabulary systems [32], while for limited vocabularies word-based models can be used. Triphone models model the possible variants of a single phoneme with different preceding and following phonemes separately. The internal structure of a phoneme is typically modeled as a chain of 3 to 5 HMM states, with the only allowed state transitions being back to the current state and forward to the next state in the chain [13], as seen in figure 2.2.

As the modeled speech features are continuous, the observation probabilities of the HMM states are described by probability density functions. Most commonly a mixture of Gaussian distributions is used to model the speech feature space. For a Gaussian mixture model (GMM), the observation probabilities are defined as [38]

$$b_i(\mathbf{o}) = \sum_{k=1}^{K} c_{ik}\, \mathcal{N}(\mathbf{o}|\boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik}), \tag{2.3}$$

where $\boldsymbol{\mu}_{ik}$ and $\boldsymbol{\Sigma}_{ik}$ are the mean vector and covariance matrix for the kth Gaussian distribution of the observation probabilities for state S_i. Separate duration models can also be associated with the HMM states [25].

To use a trained HMM in speech recognition, the likelihood that the model produced a given sequence of observations, given the known state transition and observation probabilities, needs to be estimated. It is not possible to simply select the states that individually are most likely to have produced the observations, as the resulting state sequence can contain state transitions that are not allowed, with a transition probability a_{ij} of zero. The Viterbi algorithm [45] can be used to find the most likely path, that is, the state sequence that optimizes the joint probability of the state transitions and observations. The path probability can then be used as the likelihood of the HMM to produce the observed features.

In training the HMM, the state sequence and the sequence of observations $\mathbf{O}$ are known, and the HMM model state transition probabilities a_{ij} and observation probabilities $b_i(\mathbf{o})$ need to be estimated. Denoting the HMM parameters with λ, maximizing the probability $P(\mathbf{O}|\lambda)$, the likelihood of the training observations, leads to a model that produces high likelihood scores for the training data. No analytical optimal solution for the parameters is known, but iterative solutions are possible. In the Baum-Welch algorithm, a transformation $\mathscr{T}$ presented in [3] with the property that $P(\mathbf{O}|\mathscr{T}(\lambda)) \geq P(\mathbf{O}|\lambda)$ for any HMM parameters λ is used with an iterative expectation-maximization (EM) algorithm to arrive at a reasonable solution for the model parameters.

The HMM approach has achieved wide popularity in speech recognition, even though the Markov assumption is unrealistic for speech signals [14]. The basic training scheme is also based solely on increasing the likelihood of correct models, with no consideration of the scores of incorrect models. Discriminative training [14] can alleviate this problem, though a detailed description is outside the scope of this work. For limited vocabularies, template matching systems using dynamic time warping can also be used.

2.3 Language modeling

In large vocabulary speech recognition, the acoustic differences between words are not sufficient for reliable recognition. Consider the English words “flower” and “flour”, which have a similar pronunciation. Contextual information about the surrounding text can, however, often be used to find out which interpretation is more likely. For example, if we have successfully recognized the previous word as “beautiful”, the whole utterance is more likely to be “beautiful flower”. To provide this kind of insight, language models are used.

Formally, the language model can provide a prior probability $P(W)$ for the likelihood of a given word sequence $W = \{w_1, \ldots, w_K\}$. Typically the sequence probability is defined as a product of the conditional probabilities of individual words, [19]

$$P(W) = \prod_{k=1}^{K} P(w_k|w_1, \ldots, w_{k-1}), \tag{2.4}$$

where the probability of each word depends on the preceding words in the sequence. The available context is shorter at the beginning of the sequence, and usually a special “start of sentence” token is explicitly added as the first word to provide some context for the first actual content word of the sequence.

N-gram models are a class of language models commonly used in LVCSR systems. The assumption made by an n-gram model of order N is that the probability of a word depends only on the $N - 1$ preceding words, [19]

$$P(W) = \prod_{k=1}^{K} P(w_k|w_{k-N+1}, \ldots, w_{k-1}). \tag{2.5}$$

In general, the probabilities are estimated based on the relative frequencies of particular N word sequences in a large training corpus of textual data.

To model far-reaching dependencies, long n-grams have to be used. However, for large vocabularies the amount of possible n-grams rises rapidly as the order of n-grams is increased, and it is not possible to provide a training corpus containing examples of all valid n-grams. Approaches for model smoothing, such as back-off and interpolation, can be used to lessen the impact of the sparsity of the training data [6].

For agglutinative languages such as Finnish, the number of possible inflections makes it difficult to model the language successfully with a word-based model. Sub-word units can be used for better model coverage, and the use of units resembling linguistic morphemes is especially effective [17]. An unsupervised method called Morfessor for extracting a lexicon of morpheme-like units from a text corpus was presented in [7]. The shorter unit size, however, requires the use of longer n-grams than with a word-based model in order to capture an equally long context. Model pruning or growing methods can be used to build variable-length n-gram models that include frequently occurring longer n-grams without needlessly expanding the model size [43].

2.4 Decoding

The task of the decoder is to find the optimal word sequence $\hat{W}$ of Equation (2.1) in a computationally efficient way. A number of different solutions to the problem have been used in LVCSR systems [2].

The information provided by the acoustic and language models can be combined into a search network [19]. The HMM chains corresponding to individual phonemes can be combined to form models of entire words, and the word network themselves combined using the language model probabilities to form the complete utterance. The decoding task then is to find the word sequence corresponding to the best path through the search network.

The search network can be represented as a tree structure, where alternative recognition hypotheses starting with the same prefix share the same tree branch. In large vocabulary speech recognition, the tree of possible hypotheses is extremely large, and needs to be built up dynamically as needed. Efficient pruning of unlikely hypotheses from the search tree is also required for acceptable performance. For the one-pass Viterbi beam search decoder used in this work, the pruning is performed by discarding at each time step all paths whose likelihood differs more than a given threshold from the current best path. Alternative approaches to the decoding problem are certainly possible and can be found for example in [2].

The decoder used in this work utilizes a single-pass time-synchronous Viterbi beam search algorithm [36]. The use of sub-word morpheme-like units in the language model affects the design of the decoder, as word breaks need to be modeled explicitly and the search space depends on the way words are split into morphs [18].

Chapter 3

Spectrum estimation methods in speech feature extraction

The theoretical background for the spectral envelope estimation methods investigated in this thesis is presented in this chapter.

The conventional MFCC feature extraction process is described in detail in Section 3.2. Section 3.3 presents the mathematical background for linear prediction necessary for the discussion on perceptual linear prediction in Section 3.4 and weighted linear prediction in Section 3.5. Finally, the alternative minimum variance distortionless response (MVDR) spectral envelope estimation method is introduced in Section 3.7.

3.1 Spectral envelope estimation

Consider the following simplified model for voiced speech production, where the speech signal $s(n)$ is formed as the convolution [9]

$$s(n) = e(n) \, * \, \theta(n), \tag{3.1}$$

where $e(n)$ is the excitation source and $\theta(n)$ the vocal tract response. For speech recognition, extracting the vocal tract response and discarding the excitation information from the resulting signal is useful, as the information relevant for distinguishing the spoken words is mainly in the vocal tract response, while the excitation source primarily contains the irrelevant pitch information.

In the spectral domain, the convolution of Equation (3.1) is represented as the multiplication of the two spectra. In this model, the excitation source resembles a periodic pulse train, causing the comb-like harmonic structure of the resulting spectrum, while the vocal tract resonances provide the contour of the smooth upper envelope of the spectrum [14]. It therefore follows that modeling the upper envelope amounts to recovering the vocal tract information, while discarding the excitation source.

A typical spectrum estimated with the FFT algorithm for a voiced speech waveform is shown in Figure 3.1, along with a spectral envelope estimate. The strongest resonances of the vocal tract are called *formants*, and they are the primary means of distinguishing between vowels. The formant frequencies can be seen in the figure as peaks in the spectral envelope.

In this work, spectral envelope estimation methods based on linear prediction (sections 3.3, 3.4, 3.5) and minimum variance distortionless response modeling (Section 3.7) are considered. The cepstral smoothing performed by the mel-frequency cepstral coefficient feature extraction process (Section 3.2) can also be seen as a spectral envelope estimation method. Alternative methods

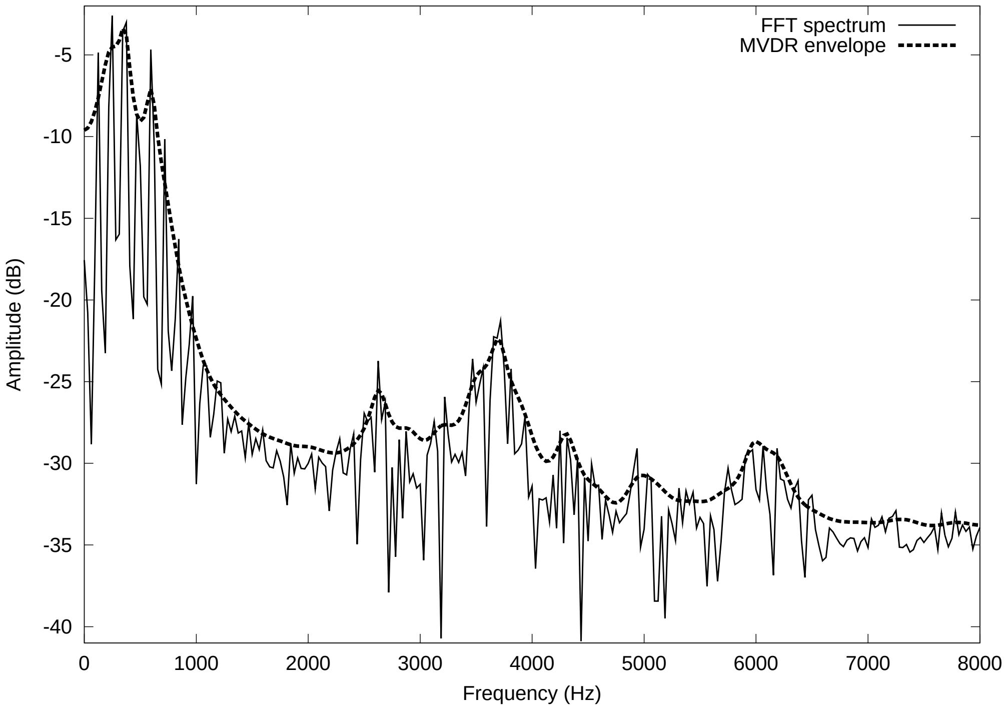

Figure 3.1: *Speech spectrum for phoneme /u/*

are presented in [41], among others.

3.2 Mel-frequency cepstral coefficient features

Mel-frequency cepstral coefficients (MFCC) are a standard feature representation used for speech recognition purposes. It has several advantages over simpler solutions such as the linear frequency cepstrum. As it is based on the psychoacoustic mel scale, it approximates the frequency-sensitive spectral resolution of human hearing, suppressing insignificant changes in high-frequency bands of the spectrum. A reasonably small number of mel-frequency cepstral coefficients also capture well the phonetically significant information, while discarding speaker-dependent variation. [8]

Cepstral processing of speech is motivated by the voiced speech production model presented in Section 3.1. According to the model, the spectrum of speech is formed as the multiplication of the spectra of the excitation source and the vocal tract response. Correspondingly, the logarithmic spectrum is simply the sum of the two components. In the cepstrum, which is obtained as the inverse Fourier transform of the log-magnitude spectrum, the two components are still linearly combined. However, as the excitation spectrum is characterized by pulse-like fast variation in the spectral values, while the vocal tract response is responsible for the smooth envelope of the spectrum [9], the Fourier transform separates the components, as the vocal tract response information is encoded mainly by the low-index values of the cepstral "quefrency" axis, while the excitation information is in the "high-quefrency" area.

Mel scale is a perceptual, nonlinear frequency scale corresponding to the *perceived* frequency, as defined by the following measurement [9, 24]. Given an arbitrary reference frequency of 1 kHz, designated as 1000 mel units, human listeners were asked to alter it so that it was a given fraction or multiple of the original frequency. The resulting physical frequencies where then labeled with the corresponding fraction or multiple of 1000 mels. The frequency scale constructed by this experiment is approximately linear below 1 kHz and logarithmic above. Various approximations of the mel scale have been used in speech recognition systems, either by using explicit linear spacing below a threshold frequency and logarithmic above it, or a simpler logarithmic approximation.

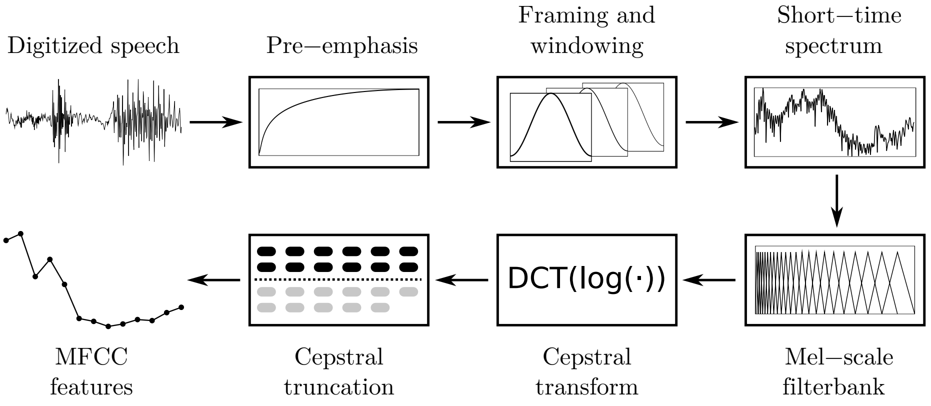

Figure 3.2: *The MFCC feature extraction process*

The MFCC feature extraction process is shown in Figure 3.2. After an initial high-pass pre-emphasis filter, short-time spectrum analysis is done by windowing the signal into frames. The linear spectrum is converted to the logarithmic mel-scale spectrum, and the discrete cosine transform (DCT) is performed to derive the cepstral coefficients. Finally, the cepstral coefficient vectors are truncated to the desired length.

The mel spectrum calculation is based on an estimate of the short-time spectrum of the analyzed audio signal. For conventional MFCC features, this estimate is obtained as the magnitude of the Fourier transformation of a short time window. However, we can easily construct other MFCC-based feature extraction systems by simply replacing the spectrum estimate with a suitable alternative. Most of the systems evaluated in chapter 4 fall into this category.

A pre-emphasis high-pass filter is often applied to the audio signal before any further processing. Typically a first-order finite impulse response (FIR) filter of the form

$$H(z) = 1 - \alpha z^{-1} \tag{3.2}$$

is used, with the α parameter commonly having values in the range from 0.9 to 1 [9]. The pre-emphasis filtering removes any DC offsets from the speech signal [14] and equalizes the typically downward slope of the voiced speech spectrum at high frequencies [19], in order to give equal importance to all spectral regions. The typical shape of a voiced speech spectrum can be seen for example in Figure 3.1.

To obtain estimates of the short-time spectrum, the audio signal is divided into short, overlapping frames. Chosen frame length varies from 10 to 25 milliseconds, and the amount of overlap from 50 % upwards. To avoid artifacts caused by discontinuities at the edges of the frame, the frames are windowed with a suitable function, such as the Hamming window function. [14, 39]

In standard MFCC processing, the spectrum estimate is based on the Fourier transformation. Combined with the windowing, we get the short-time discrete Fourier spectrum $S(k, \tau)$ of frame τ as [33]

$$S(k, \tau) = \sum_{n=0}^{N-1} w(n)\, s(n, \tau)\, e^{-jkn\frac{2\pi}{N}}, \tag{3.3}$$

where $w(n)$ is the window function, $s(n, \tau)$ the unwindowed audio signal portion for frame τ, and N the frame length in samples. In real applications, the Fourier transform is generally computed with the more computationally efficient FFT algorithm.

The frequency-dependent spectral resolution [14] of the human hearing is approximated by the MFCC representation using a mel-filterbank built of logarithmically spaced triangular bandpass filters, applied to the bins of the discrete spectrum $S(k, \tau)$. The output energies of the filters form the mel-scale spectral representation. As the differences in the detailed definition of the mel scale are not significant for speech recognition purposes [8], in this work we use a filterbank of 23 logarithmically spaced filters, with center frequency f_i (in Hz) of the ith filter given by

$$f_i = \frac{1400}{2} \left[\left(1 + \frac{R}{1400} \right)^{i/24} - 1 \right], \qquad i \in [1, 23] \tag{3.4}$$

where R is the sampling rate (in Hz). The filterbank for the 16 kHz sampling rate is shown in Figure 3.3.

The discrete cosine transform (DCT) is used to convert the logarithm of the mel-scale spectrum to a cepstral form. For a given frame τ, denoting the output energy of the kth filterbank output by $M_{k,\tau}$, we get the ith cepstral coefficient $c(i, \tau)$ with [8]

$$c(i, \tau) = \sum_{k=1}^{23} \log |M_{k,\tau}| \cos \left(i\,(k - 0.5)\,\frac{\pi}{23} \right) \tag{3.5}$$

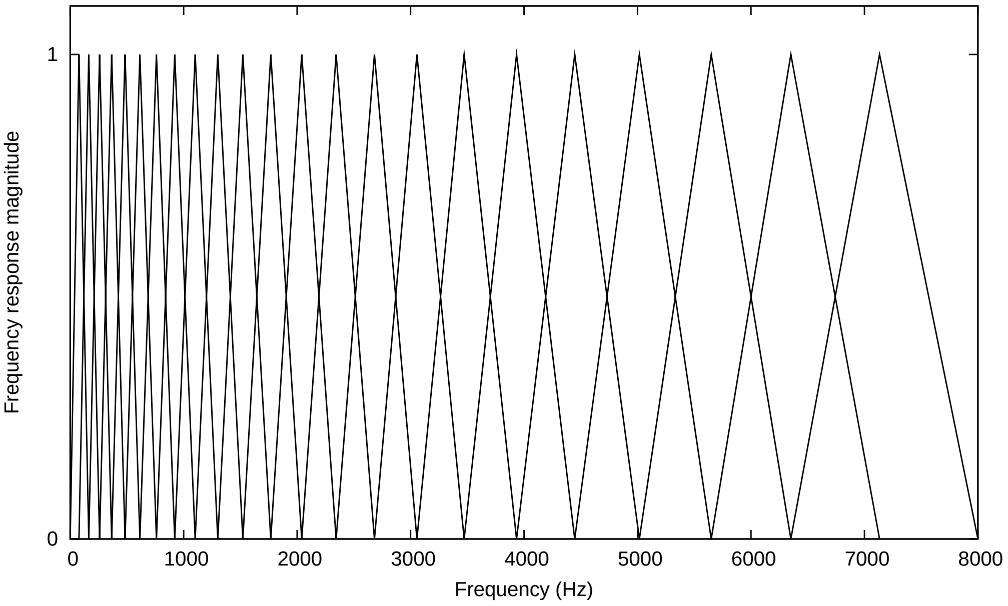

Figure 3.3: *The mel-scale filterbank used in MFCC computation*

Typically approximately twelve [19] first cepstral coefficients are used in the feature vector, as the higher-order coefficients mainly encode speaker-dependent information [8]. Application of the DCT to the mel spectrum results in nearly decorrelated cepstral values. This makes it possible to model them using Gaussian distributions with diagonal covariance matrices, which makes the task of acoustic modeling with hidden Markov models less complex [19].

Further advantages to using the MFCC features instead of a plain short-time magnitude spectrum include the possibility for cepstral mean subtraction (CMS) [14]. Convolutional degradation, for example from the signal recording path, corresponds to multiplication in the spectral domain, and furthermore to a sum in the log-spectral domain. As the cepstral coefficients are based on the log-spectrum values, subtracting a long-term average from the cepstral speech features can therefore remove time-invariant signal degradation, leaving the original speech signal intact.

In addition to the cepstral coefficients, the logarithmic frame energy is often used in the final speech features. For conventional MFCC, this corresponds to the 0th cepstral coefficient, which can be seen by setting $i = 0$ in Equation (3.5). To capture the dynamic behavior of the speech signal, the so-called delta and delta-delta features [12] are also appended to the feature vector.

These are estimates for the first and second order derivatives of the cepstral coefficients over neighboring frames, and capture information about the correlation of the static features between frames. Given a delta-window length L (in our experiments $L = 2$) in frames, the dynamic features are calculated as

$$\Delta(i, \tau) = \sum_{k=1}^{L} k \left(c(i, \tau + k) - c(i, \tau - k) \right) / K, \tag{3.6}$$

$$\Delta\Delta(i, \tau) = \sum_{k=1}^{L} k \left(\Delta(i, \tau + k) - \Delta(i, \tau - k) \right) / K, \tag{3.7}$$

$$K = \left(4L^3 + 6L^2 + 2L \right) / 6. \tag{3.8}$$

The MFCC feature extraction process described in this section can be used in combination with the spectral envelope estimation methods by replacing the Fourier spectrum estimate of Equation (3.3) with an envelope estimate. The LP-MFCC features use the conventional linear prediction models of Section 3.3, while the WLP-MFCC and SWLP-MFCC features use the unstabilized and stabilized variants of the weighted linear prediction algorithm described in Section 3.5, respectively. Finally, the MVDR-MFCC features are based on the minimum variance distortionless response modeling discussed in Section 3.7.

3.3 Linear prediction

Linear prediction is a traditional signal processing tool for constructing parametric all-pole statistical models of arbitrary signals. In the context of speech processing, linear prediction is used to derive a parametric representation for the spectral envelope of a speech signal. The model produced by linear prediction,

$$H(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}}, \tag{3.9}$$

is an all-pole model. It models non-nasal voiced sounds well, but for nasals and fricatives the effect of the zeros involved in the detailed acoustic model of speech production have to be modeled using a sufficient number of poles. The spectral resolution of the LP model depends on the model order parameter. As the model order is increased, the LP model tracks the spectrum more closely, but using too large number of poles causes the model to start tracking the harmonic structure of speech, as can be seen in the comparison of LP and minimum variance distortionless response (MVDR) models of high order in Figure 3.8.

The parametric representation of the spectral envelope has many potential uses. For example, since a relatively small number of parameters can be used to describe well the salient features of the analyzed sound, linear prediction can be used in audio coding applications to encode a speech signal on a channel with a very low bit rate [9]. In the context of this thesis, the spectral envelope models provided by linear prediction or its variants can be used to derive speech features that perform better, especially in noisy conditions.

In linear predictive signal modeling, a sample x_n is estimated by a linear combination of p previous samples, [29]

$$\hat{x}_n = -\sum_{i=1}^{p} a_i x_{n-i}, \tag{3.10}$$

where $a_i \in \mathbb{R}$ are the linear prediction coefficients for that particular signal, and p is called the *model order.* By denoting $\mathbf{a} = [1\ a_1\ \cdots\ a_p]^T$ and $\mathbf{x}_n = [x_n\ x_{n-1}\ \cdots\ x_{n-p}]^T$ we can represent the prediction error ε_n in matrix form as

$$\varepsilon_n(\mathbf{a}) = x_n - \hat{x}_n = \mathbf{a}^T\mathbf{x}_n. \tag{3.11}$$

The optimal linear prediction coefficient vector $\mathbf{a}$ for a given signal is obtained by using the squared prediction error as a cost function $\mathscr{E}(\mathbf{a})$ to minimize. The cost function written in matrix form is

$$\mathscr{E}(\mathbf{a}) = \sum_{n=1}^{N+p} (x_n - \hat{x}_n)^2 = \mathbf{a}^T\mathbf{R}\mathbf{a}, \tag{3.12}$$

where N is the frame length, and $\mathbf{R} = (r_{i,j})$ is the $(p+1) \times (p+1)$ autocorrelation matrix,

$$\mathbf{R} = \sum_{n=1}^{N+p} \mathbf{x}_n\mathbf{x}_n^T. \tag{3.13}$$

Here the signal x_n is assumed to be zero when n is outside the range $[1, N]$, leading to the *autocorrelation method* of LP coefficient estimation, for which the resulting model is known to always be stable [29].

The minimization problem

$$\min_{\mathbf{a}}\ \mathscr{E}(a) = \min_{\mathbf{a}}\ \mathbf{a}^T\mathbf{R}\mathbf{a} \tag{3.14}$$

can be solved by setting $\frac{\partial \mathscr{E}}{\partial a_i} = 0, i \in [p]$. This condition results in the Yule-Walker system of p equations,

$$\begin{bmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,p} \\ r_{2,1} & r_{2,2} & \cdots & r_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p,1} & r_{p,2} & \cdots & r_{p,p} \end{bmatrix} \mathbf{a} = - \begin{bmatrix} r_{1,2} \\ r_{1,3} \\ \vdots \\ r_{1,p+1} \end{bmatrix}. \tag{3.15}$$

The autocorrelation matrix $\mathbf{R}$ in Equation (3.15) is a Toeplitz matrix, and therefore the LP coefficients can be solved in a very computationally efficient way with the Levinson-Durbin recursion [29].

The LP all-pole model of Equation (3.9) can be used directly as a spectral envelope estimate in speech recognition. It is well-known, however, that the LP models do not model high pitch voiced speech well, because the harmonic frequencies are so sparsely spaced that even a low order LP model tracks the harmonic peaks instead of the spectral envelope [11]. In addition, the LP model is not especially robust to noise [31].

In this work, the LP spectral envelope estimates have been used with the MFCC feature extraction scheme discussed in Section 3.2 to generate features for speech recognition. In LP-MFCC feature extraction, after the pre-emphasis filtering and framing, the linear prediction coefficients are determined for each frame. The impulse response generated by the LP all-pole model $H(z)$ of Equation (3.9) is then used as the input signal for the rest of the MFCC processing, starting with the FFT short-time spectrum analysis, which in this case extracts the spectrum of the LP model.

3.4 Perceptual linear prediction

The conventional LP estimate follows the original spectrum uniformly over all frequencies. However, the spectral resolution of the human auditory system is markedly lower for high frequencies. In addition, the sensitivity of hearing depends on the frequency, being highest in the middle of the audible spectral range. Therefore, the sound power required for equally perceived loudness is generally speaking lower for the middle frequencies. Accordingly, the perceptual linear prediction (PLP) method proposes various changes to conventional LP modeling that approximate the irregularities of human hearing. [15]

The major difference between PLP and conventional LP is the perceptual scaling of the linear autocorrelation estimate used in conventional LP. As in conventional linear prediction, the autocorrelation estimates are computed via an inverse Fourier transformation of the signal power spectrum estimate. However, in PLP a modified spectrum on a perceptual Bark scale is used.

The Bark frequency scale is based on psychoacoustic measurements related to the masking effect in the human hearing [19, 24]. The critical bandwidth is defined as the width of the frequency range that can contribute to the masking of a pure tone at the center of the band. The Bark scale is defined so that a difference of 1 Bark corresponds to the critical bandwidth over the entire range of audible frequencies. The Bark scale closely corresponds to the mel scale discussed in Section 3.2.

Like the mel-scale spectrum, the spectrum estimate used in PLP is computed using a filterbank approximating the processes of the human hearing, derived from psychoacoustic measurements. The PLP filters are spaced at regular intervals in the Bark scale, where the Bark frequency $\Omega(\omega)$ corresponding to the angular frequency ω is given by [15]

$$\Omega(\omega) = 6 \ln \left\{ \frac{\omega}{1200\pi} + \left[\left(\frac{\omega}{1200\pi} \right)^2 + 1 \right]^{0.5} \right\}. \tag{3.16}$$

The triangular band-pass filters of the mel spectrum are replaced by filters that approximate the critical band masking behavior [14], with the frequency response in the Bark scale of a filter centered at Bark frequency Ω is [15]

$$\Psi(\Omega) = \begin{cases} 0 & \text{when } \Omega < -1.3 \\ 10^{2.5(\Omega+0.5)} & \text{when } -1.3 \leq \Omega \leq -0.5 \\ 1 & \text{when } -0.5 < \Omega < 0.5 \\ 10^{-1.0(\Omega-0.5)} & \text{when } 0.5 \leq \Omega \leq 2.5 \\ 0 & \text{when } \Omega > 2.5 \end{cases} . \tag{3.17}$$

Finally, the filterbank outputs are scaled with an approximation of the equal-loudness curve of the human hearing [19] given by the function [15]

$$E(\omega) = \frac{(\omega^2 + 5.68 \cdot 10^6)\omega^4}{(\omega^2 + 6.3 \cdot 10^6)^2 \cdot (\omega^2 + 0.38 \cdot 10^9)}. \tag{3.18}$$

This step has no direct analogy in the mel spectrum computation.

In practical implementations, each sample of the PLP spectrum is computed by applying a single weighted summation to the discrete short-time FFT

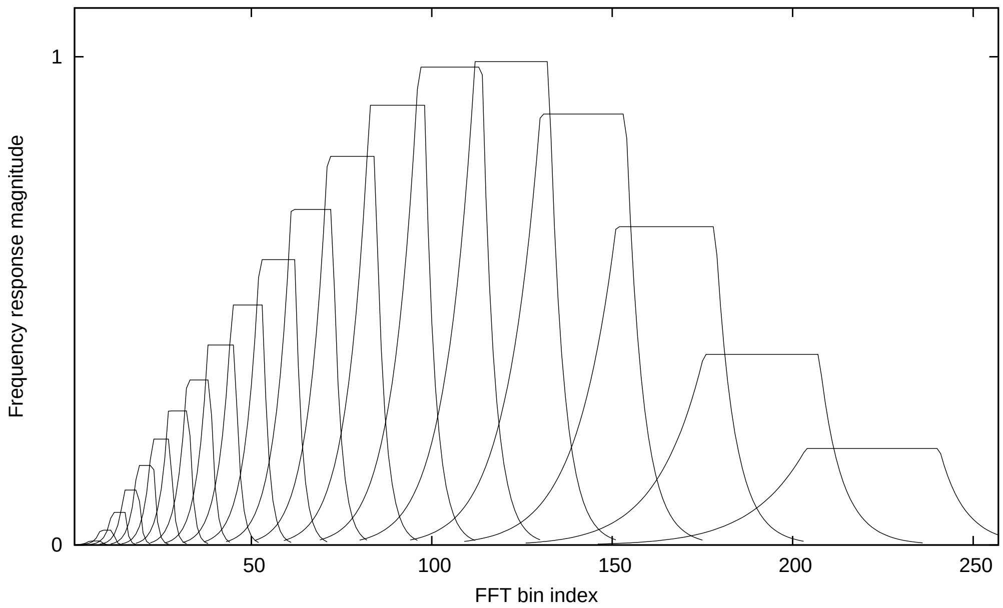

Figure 3.4: *The perceptual linear prediction filterbank. FFT bin index 257 corresponds to the Nyquist frequency of 8 kHz.*

magnitude spectrum [15]. The sum weights are chosen to implement the Bark filterbank processing and the equal-loudness scaling factor for that particular PLP spectrum sample. Figure 3.4 shows the filter shapes computed for a 512-point FFT, using 21 PLP filters designed to analyze a signal with a 16 kHz sampling rate.

The non-linear compression in the human auditory system [19], modeled in the MFCC computation with the use of the logarithmic spectrum values, is dealt with in PLP feature extraction by taking the cubic root of the PLP power spectrum, as the cubic root approximates better the relationship between sound intensity and perceived loudness [15]. This is the final step before the autoregressive processing of the "perceptual autocorrelation" values obtained with the inverse discrete Fourier transform from the final PLP spectrum. The linear predictive modeling in other respects follows the conventional LP method given in section 3.3.

3.5 Weighted linear prediction

Weighted linear prediction [26] is based on applying a temporal weight term to the conventional LP cost function of Equation (3.12). The temporal weight function can therefore be used to guide the LP algorithm to focus on modeling particular temporal regions of the input signal. In the case of noisy speech, for example, more importance can be given to high-energy temporal regions, where the relative effect of the noise is less prominent. In practice, the weighting is implemented by replacing the autocorrelation matrix $\mathbf{R}$ of Equation (3.13) with a weighted autocorrelation matrix of the form

$$\mathbf{R} = \sum_{n=1}^{N+p} w_n \mathbf{x}_n \mathbf{x}_n^T, \tag{3.19}$$

where $w_n \geq 0$ are the temporal weights and N, p are again the frame length and model order, respectively.

The weighted linear prediction coefficient vector $\mathbf{a}$ can be solved from the LP minimization problem of Equation (3.14) using the autocorrelation matrix from Equation (3.19). In this case, however, the symmetric autocorrelation matrix does not necessarily have the Toepliz structure, so the Levinson-Durbin recursion cannot be used to obtain the solution, and more complex methods have to be used [5, 28].

The linear predictive algorithm as described in section 3.3 guarantees the stability of the resulting all-pole model [39]. No such guarantees are possible for the basic weighted linear prediction approach with an arbitrary weight function, however. As the stability of the model is a desired property in many applications, especially in the speech coding and synthesis areas, the stabilised weighted linear prediction (SWLP) [27, 28] method modifies the weighted autocorrelation matrix $\mathbf{R}$ as follows to ensure the stability of the resulting system.

It is possible to express the matrix $\mathbf{R}$ of Equation (3.19) as $\mathbf{R} = \mathbf{Y}^T\mathbf{Y}$, where

$$\mathbf{Y} = \begin{bmatrix} \sqrt{w_1}x_1 & 0 & & \vdots \\ \sqrt{w_2}x_2 & \sqrt{w_2}x_1 & & \\ & \sqrt{w_3}x_2 & & 0 \\ \vdots & & & \sqrt{w_{p+1}}x_1 \\ \sqrt{w_N}x_N & \vdots & \ddots & \sqrt{w_{p+2}}x_2 \\ 0 & \sqrt{w_{N+1}}x_N & & \\ & 0 & & \vdots \\ \vdots & & & \\ & \vdots & & \sqrt{w_{N+p}}x_N \end{bmatrix}, \tag{3.20}$$

and the columns $\mathbf{y}_k$ of matrix $\mathbf{Y}$ can be recursively constructed as

$$\begin{aligned} \mathbf{y}_1 &= [\sqrt{w_1}x_1 \ \cdots \ \sqrt{w_N}x_N \ 0 \ \cdots \ 0]^T, \\ \mathbf{y}_k &= \mathbf{B}\,\mathbf{y}_{k-1}, \qquad k = 2, 3, \ldots, p+1, \end{aligned} \tag{3.21}$$

using the matrix

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ \sqrt{w_2/w_1} & 0 & 0 & \cdots & 0 \\ 0 & \sqrt{w_3/w_2} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \sqrt{w_{N+p}/w_{N+p-1}} & 0 \end{bmatrix}. \tag{3.22}$$

The SWLP method produces a model guaranteed to be stable by altering the nonzero elements of $\mathbf{B}$ to be [27]

$$\mathbf{B}_{i+1,i} = \begin{cases} \sqrt{w_{i+1}/w_i}, & \text{if } w_i \leq w_{i+1} \\ 1, & \text{if } w_i > w_{i+1} \end{cases} \tag{3.23}$$

It can be shown [28] that this leads to a stable model.

As was mentioned earlier in this section, in some applications such as audio coding the stability of the model is essential. In the speech recognition feature extraction task considered in this thesis, however, only a spectrum estimate represented by the model is needed. We therefore have to extract the frequency response of the LP *synthesis filter* $H(z)$ of Equation (3.9) for the obtained WLP coefficients, as that is the spectral envelope estimate used when computing the MFCC features.

The LP *analysis filter* $A(z)$ is the inverse of the synthesis filter,

$$A(z) = \frac{1}{H(z)} = 1 - \sum_{k=1}^{p} a_k z^{-k}. \tag{3.24}$$

The analysis filter is a FIR filter and therefore stable, and its magnitude spectrum can be computed via FFT of the filter coefficients [44]. The magnitude response of the synthesis filter is the inverse of this spectrum [35]. As the model is not guaranteed to be stable, the magnitude spectrum of the analysis filter can be zero, which would lead to infinite values in the final spectral envelope estimate. To avoid this, all values of the magnitude spectrum of the analysis filter that are more than a given amount below the maximum can be clamped to that value. In this work, a threshold value of 80 dB below the maximum has been used, as it is well above the dynamic range of the input signal.

The stabilization step causes some further, uncontrollable smoothing of the spectral envelope estimate, as can be seen in Figure 3.5, where WLP and SWLP models of the same order are compared. The formant frequencies important for distinguishing voiced sounds are clearly more pronounced in the WLP estimate.

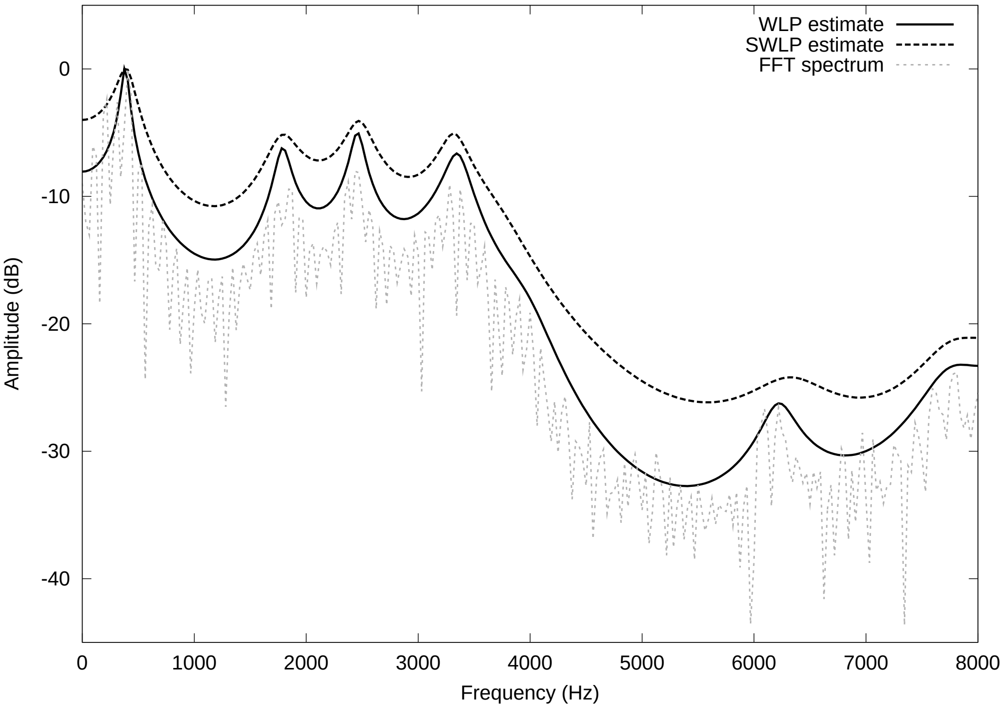

Figure 3.5: *Unstabilized and stabilized WLP envelope estimates of a voiced clean speech frame, with the same model order of 20 and STE window width of 16 samples*

The choice of the temporal weights w_i strongly affects the models generated by weighted linear prediction. For various reasons, use of the short-time energy (STE) function as the weight function has been suggested [26, 28] for speech processing tasks. The STE weight w_n for the nth sample x_n is computed by taking the signal energy of a sliding window of M previous samples,

$$w_n = \sum_{i=n-M}^{n-1} x_i^2. \tag{3.25}$$

Using the STE weight function causes the WLP model to emphasize highly energetic regions of the speech frame. In these regions, the noise energy relative to the signal energy is lower, especially in case of stationary additive noise. In addition, for a voiced speech frame the peaks of the STE function coincide with the closed phase of the glottal cycle, when the formants are also most easily distinguished [28]. The short-time energy function for a short sample of voiced speech can be seen in Figure 3.6.

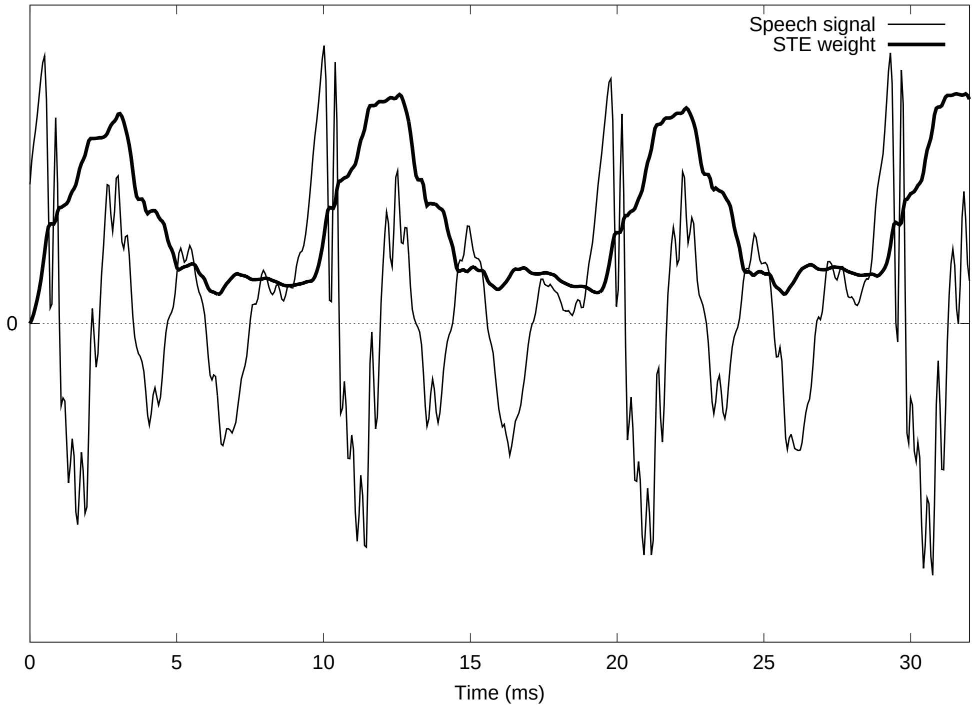

Figure 3.6: *STE weight function calculated with a window width of 3 ms (48 samples) for a 32 ms sample of voiced phoneme /i/*

The short-time energy window width parameter M can also be controlled to tune the spectral estimates represented by the weighted linear prediction models. As the M parameter is increased, the size of the portion of the frame covered in Equation (3.25) grows, causing the variance between weights of successive samples to decrease. From the WLP autocorrelation matrix of Equation (3.19) it can be seen that if the sample weights are set to a constant value, the generated all-pole model exactly matches the conventional linear predictive model. Correspondingly, increasing the M parameter causes the WLP or SWLP modeling behavior approach that of conventional LP, with the more pronounced spectral peaks. In contrast, using a shorter STE window results in more smoothed spectral estimates, as the model emphasizes more strongly short sections of the speech frame. This behavior can be observed in the spectral envelope examples of Figure 3.7.

The M parameter can be given a fixed value based on experiments on development set data, like is commonly done with the model order parameter. This approach has been taken in the experiments of sections 4.4, 4.5 and

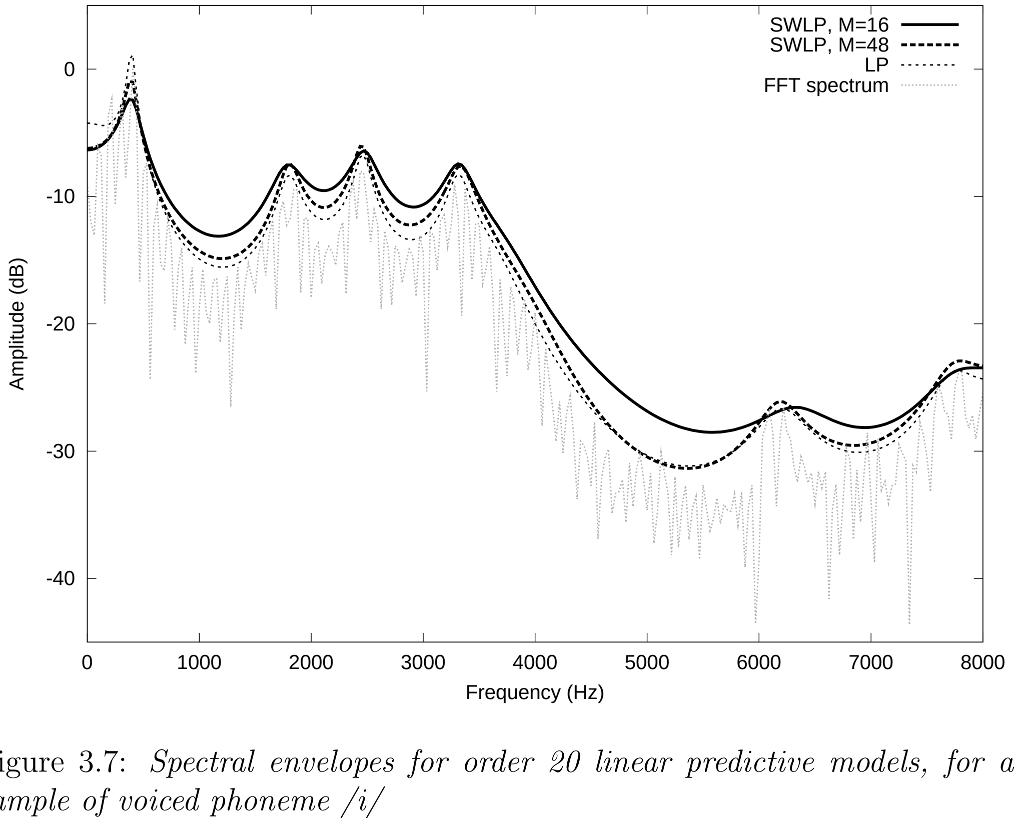

Figure 3.7: *Spectral envelopes for order 20 linear predictive models, for a sample of voiced phoneme /i/*

4.6. However, the optimal M value seems [23, 35] to be different for example for different amounts of environmental noise in the speech data being recognized. This naturally prompts the question whether the M value could be adaptively updated during the recognition, based on changes in the incoming audio data. Some promising preliminary results along these lines have been seen in [23, 35] and the methods described in Section 3.6 have been experimentally evaluated in sections 4.7 and 4.8 of this thesis.

3.6 Weight function parameter adaptation

Performance of the SWLP feature extraction method depends on the correct selection of the STE weight function window width M. Instead of using a constant value for M, it is possible to make an adaptive system where the weight function window width used in the analysis of the τth frame of the ith utterance of the test set is a function $M(i, \tau)$. The $M(i, \tau)$ value can depend on any suitable characteristics, such as for example per-frame SNR estimates from on-line noise models, or confidence measures based on likelihood values from the models used by the recognizer.

In the adaptation methods described in this section, we consider only the task of selecting separate M values on a per-utterance level, so $M(i, \tau) = M(i)$. There is no theoretical reason why the M parameter could not be updated for each frame τ separately, and the limitation is simply to make it easier to evaluate the possible methods, as their effect can then be observed without having to implement any changes to the speech recognition system. The final recognition results for any per-utterance adaptation method can be constructed by first recognizing each utterance with an array of fixed M values, and then selecting the per-utterance results corresponding to the adapted $M(i)$ values.

For each of the N utterances in a test set, the parameter adaptation method has to select a suitable window width parameter value $M(i)$, where i is the index of the utterance. We want to optimize the average recognition error rate E for the entire test set,

$$E = \frac{\sum_{i=1}^{N} L_i E_i^{M(i)}}{\sum_{i=1}^{N} L_i}, \tag{3.26}$$

where E_i^M is the error rate for the ith individual utterance when using window width M in the recognition, and L_i is the length of the utterance, measured in letters or words, depending on the type of the error rate in question. The term *error rate* is here used to mean either the word or letter error rate as described in Section 4.3.

3.6.1 Oracle-based M adaptation

In order to find a theoretical lower bound for E, knowledge of the correct recognition result can be used. As the E_i^M values are then known, the optimal $M(i)$ for equation (3.26) can simply be selected as

$$M(i) = \arg\min_M E_i^M. \tag{3.27}$$

This method is in this work referred to as oracle-based M adaptation. Experimental results obtained using it are presented in Section 4.7.

3.6.2 M adaptation based on acoustic model probabilities

In a practical adaptation method, the exact E_i^M values in equation (3.26) cannot be known. However, different confidence measures can be used to give an estimate for the reliability of the recognition result [21]. The acoustic model probability based M adaptation uses the per-frame observation probabilities of the acoustic HMM described in Section 2.2.

A given utterance is first recognized using an array of pre-defined fixed M values. As described in Section 2.4, the speech recognition hypothesis is formed as a path through a search network of the acoustic HMM states. The states of the winning path corresponding to the selected recognition hypothesis are here denoted by $q_{i,\tau}^M$, where i is the utterance index, τ the specific speech frame and M the fixed window width used in the recognition.

For any observed speech feature vector $\mathbf{o}$ and acoustic HMM state q, the observation probabilities of the HMM can be used to determine the logarithmic likelihood value $a(\mathbf{o}|q)$ for observing the particular feature vector in that state. As the choice of M affects the feature extraction, the observed speech features $\mathbf{o}(i, \tau, M)$ depend on both the input frame and the M value that was used. Some of the acoustic model states describe silent parts of speech, such as breaks between words, and the set of these states is here denoted by S.

Using these notations, we define the final per-utterance window width $M(i)$ using the average per-frame acoustic likelihood of the states $q_{i,\tau}^{M}$ on the winning recognition hypothesis path, for any non-silent frames, as

$$M(i) = \arg\max_{M} \frac{1}{N_i^M} \sum_{\substack{\tau, \\ q_{i,\tau}^M \notin S}} a(\mathbf{o}(i, \tau, M) | q_{i,\tau}^M), \tag{3.28}$$

where N_i^M is the number of frames for which the winning path was in non-silent HMM states. The recognition result corresponding to the selected $M(i)$ value is then used as the final result. Results of M adaptation using this criterion are given in Section 4.8.

The average per-frame acoustic model likelihood, though not with the elimination of silent states, has been used as a basis for a confidence measure in several tasks. In [4] a similar measure was used as one of the features for word verification in a key-word spotting system, while in [42] the method was selected as one of the features for a confidence tagger for spontaneous speech. Both systems also utilized a set of other features, and a possible topic for future work is to see whether the M adaptation described here could be enhanced with an improved confidence measure.

As the M adaptation method discussed in this section is based on the acoustic HMM state paths of k final recognition hypothesis, for k different values of M, it causes a k-fold increase in computational cost of the recognition. An important topic for future work is to find effective adaptation methods that do not need the decoder output for the M selection.

3.7 Minimum variance distortionless response modeling

Minimum Variance Distortionless Response (MVDR) method for modeling the spectral envelope addresses many of the shortcomings of the conventional LP model. Notably, the MVDR spectral estimate is much less prone to modeling the sharp contours of the harmonic structure of speech, even for high pitch speech where the harmonics are located more sparsely, as can be seen in Figure 3.8. The MVDR spectrum can also be efficiently computed from conventional LP coefficients, which is useful in many applications. [31]

Conceptually, the MVDR method estimates the signal power at a given fre-

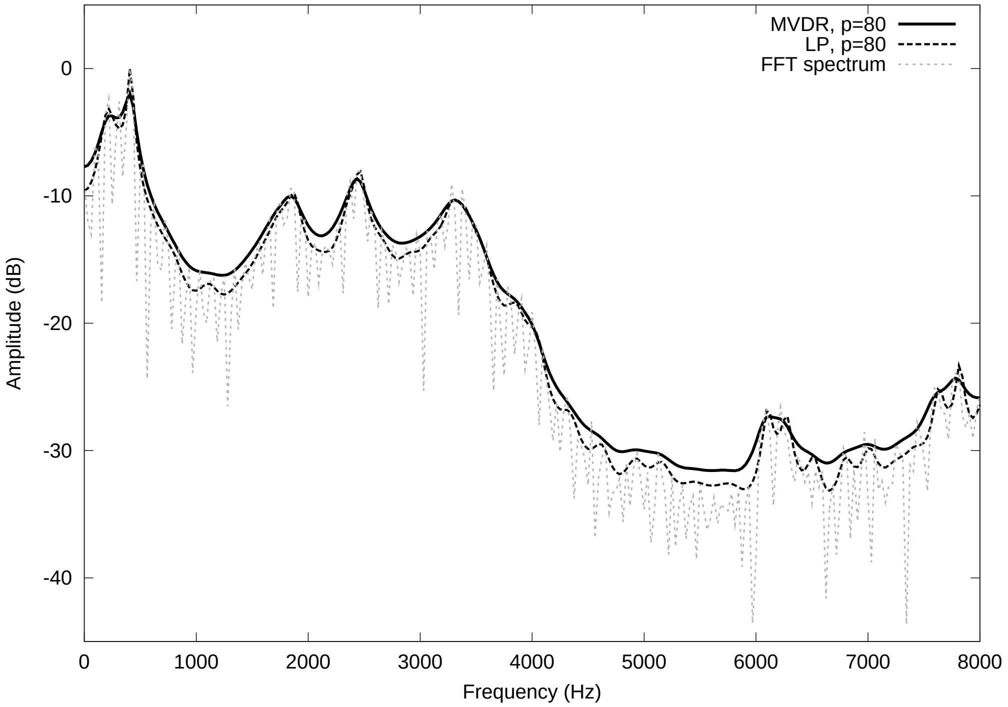

Figure 3.8: *Spectral envelope comparison for linear predictive (LP) and minimum variance distortionless response (MVDR) models of high order 80, for a sample of voiced phoneme /i/*

quency ω_l with the output power of a specific Mth order FIR filter with impulse response $h_l(n)$. This filter is designed to produce the lowest possible output power for the particular speech frame being analyzed, under the *distortionless constraint*, that it has a unit frequency response at frequency ω_l: [31]

$$H_l(e^{j\omega_l}) = \sum_{k=0}^{M} h_l(k) e^{-j\omega_l k} = 1 \tag{3.29}$$

It is possible to express also the FFT based periodogram spectrum estimate as the output of a filterbank, constructed from a fixed set of filters that are independent of the analyzed data or the center of the frequency band. In contrast, the MVDR filters are designed to be optimal for a particular speech frame and center frequency. The use of this constrained optimization filter design enhances the bias and variance properties of the MVDR spectrum estimate. Bias in the output at a specific frequency caused by leakage of power from surrounding frequencies through the band-pass filter is reduced, as each band-pass filter is specifically designed to have as small side-lobes as possible in any nearby frequency regions that have high energy. The improvement in the variance of the output comes from averaging several samples of the band-pass filter output, instead of using just a single sample as is done in the periodogram. [31]

The need to design a separate filter $h_l(n)$ for each frequency band and each frame of input data is purely conceptual, however, and it is actually possible to represent the MVDR power spectrum in a parametric form based on conventional LP coefficients. The Mth order MVDR power spectrum estimate $P_{\mathrm{MV}}^{(M)}$ can be written as [31]

$$P_{\mathrm{MV}}^{(M)}(\omega) = \frac{1}{\sum_{k=-M}^{M} \mu_k e^{-j\omega k}}, \tag{3.30}$$

where

$$\mu_k = \begin{cases} \frac{1}{P_e} \sum_{i=0}^{M-k} (M+1-k-2i) a_i a_{i+k}^*, & \text{for } k = 0, \ldots, M, \\ \mu_{-k}^*, & \text{for } k = -M, \ldots, -1, \end{cases} \tag{3.31}$$

and a_k are the Mth order LP model coefficients, while P_e is the LP prediction error variance.

The MVDR estimate can be used in speech recognition in various ways: as a direct replacement for the FFT spectrum in MFCC computation (called the MVDR-MFCC method), using perceptually modified autocorrelation estimates as in the PLP method (called the PMCC method) [10, 50] or possibly with direct warping of the FFT spectrum instead of a filterbank (called the PMVDR method) [51, 52].

In the PMCC approach, the autoregressive LP analysis from which the MVDR coefficients are derived is done using autocorrelation estimates obtained from a mel-smoothed spectrum. This procedure resembles closely the perceptual linear prediction (PLP) method described in Section 3.4, with the exception that the LP coefficients are then converted into the MVDR spectrum estimate. PMCC feature generation has two main advantages over MVDR-MFCC. As the mel-smoothed spectral samples are an average over several FFT samples, the variance of the resulting perceptual autocorrelation estimates is lower, and the results are more reliable. The dimensionality of the mel-scale spectrum is also lower than the original spectrum, and the computation of the MVDR model parameters therefore less computationally intensive. [10]

In the PMVDR method, the psychoacoustically motivated change in spectral resolution is provided using a suitable warping function directly on the MFCC spectrum, to obtain suitable perceptual autocorrelation estimates for MVDR modeling without the spectral smoothing step. One reason for the mel-scale spectral smoothing in the MFCC features is that the smoothing step discards some of the speaker and pitch-dependent information such as the harmonic structure that is not useful from a pure speech recognition perspective [9]. As the MVDR modeling process is itself capable of providing a good estimate of the spectral envelope, the spectral smoothing performed by the filterbank processing is redundant. [52]

Chapter 4

Experimental evaluation

This chapter presents the speech recognition experiments designed to evaluate the performance of the considered spectral envelope estimation methods. The experiments on automatically adapting the SWLP weight function window width parameter M are also described.

The SPEECON Finnish language corpus described in Section 4.1 was used as the source of speech data in all the experiments described in this work. The experiments were performed with the large vocabulary continuous speech recognition system developed in the Adaptive Informatics Research Centre at the Helsinki University of Technology. The design of the system is presented in Section 4.2. The evaluation metrics for speech recognition performance are detailed in Section 4.3. Finally, the remaining sections of this chapter describe the results obtained from the experiments. The initial small-scale feature evaluation results are presented in Section 4.4, the comprehensive evaluation of SWLP features against the MFCC baseline is described in Section 4.5, and the comparison of the unstabilized WLP features against the SWLP features is discussed in Section 4.6. Finally, sections 4.7 and 4.8 outline our experiences from the M parameter adaptation test.

4.1 Speech data used in the experiments

Material from the Finnish language version of the SPEECON [20] speech corpus was used exclusively in all experiments. SPEECON is an EU-funded project to collect speech data for 18 different languages, with the primary aim to promote the development of voice controlled user interfaces.

The SPEECON corpus includes realistic noisy speech data recorded in many different environments. The experiments described in this paper used noisy speech from the "car" and "public places" environments. The "car" environment recordings were done in a 5-passenger vehicle, driving on highways, country roads and in city traffic, with the speaker in the co-driver's seat. "Public places" recordings come from large rooms or halls, or open-air locations, and often have other people talking in the vicinity.

The noisy material was divided into separate training, development and evaluation subsets by a random selection of speakers, in order to get data sets with similar characteristics. For example, the car environment data contained metadata information about the type of car used and the general driving environment (city, country, highway).

For training purposes, two different data sets were constructed from the data allocated for training. The first training set consisted of approximately 21 hours of clean speech recorded with no background noise, from 293 separate speakers. The second data set was a multicondition training set of similar size, containing even amounts of clean and noisy speech, with the noisy data from both the car and public place environments. The distribution of signal-to-noise ratio (SNR) estimates taken from SPEECON transcription annotations for the two training sets is shown in Figure 4.1.

The tests with the multicondition training set were also the first experiments utilizing noisy speech training for Finnish language speech recognition with the speech recognizer used in the experiments.

A separate evaluation set was constructed for each recording channel of the two different noisy environments. Each utterance in the SPEECON corpus was recorded using four audio channels corresponding to microphones with different distances to the speaker in order to obtain a range of realistic noise conditions. Increasing the distance between the microphone and the speaker is a natural way of producing more challenging noise conditions. Audio from the first three SPEECON recording channels was used in the test sets.

The "car" environment evaluation set contained 30 phonetically rich sentences of read speech by 20 different speakers, with a total length of 57 minutes including the leading and trailing silences. Channel 0 was recorded with a headset microphone at a distance of 2–5 centimeters from the speaker's mouth. Channel 1 microphone was a lavalier microphone positioned the chin and the shoulder of the speaker. Finally, audio for channel 2 was recorded with a medium-distance microphone mounted at the car ceiling behind the rear-view mirror. Notably, for the car data set the SNR estimates for channel 1 were lower than those for channel 2, even though the microphone was positioned closer to the speaker. Spectral analysis suggests that the low SNR estimate for channel 1 was caused by high levels of noise at frequencies below 200 Hz, which are outside the spectral ranges most important for speech recognition.

The evaluation data set for the "public places" environment had a similar structure, with 30 sentences by 30 speakers, and a total length of 94 minutes. Recording equipment was mostly the same, with the exception that channel 2 was recoded with a different medium-distance microphone placed 0.5–1 meter away for the speaker. SNR estimates provided by the recording system for both evaluation sets can be found in Figure 4.2.

For the "car" environment, the SNR estimates given for channel 1 recordings (lavalier microphone, average SNR of 5 dB) are consistently lower than the SNR estimates for channel 2 (medium-distance microphone, average SNR of 8 dB), even though the speech recognition tests in all experiments produced better results for channel 1 than channel 2. Spectral analysis of the recordings revealed as a likely explanation that the low SNR estimate for channel 1 was caused by high noise levels at frequencies below 200 Hz, which however are outside the spectral areas important for speech recognition.

Smaller development sets of similar composition were used for tuning the parameters of the various algorithms. The "car" and "public places" development sets each had 30 sentences by 10 and 20 speakers, with a total length of 21 and 60 minutes, respectively. Averages of all per-utterance SNR estimates for the development sets were within 1 dB of the averages of corresponding evaluation sets.

4.2 Speech recognition system used in the experiments

All speech recognition experiments were performed with the large vocabulary continuous speech recognizer developed in the Adaptive Informatics Research Centre at the Helsinki University of Technology. The general structure of the system is described thoroughly in Chapter 2. This section gives information

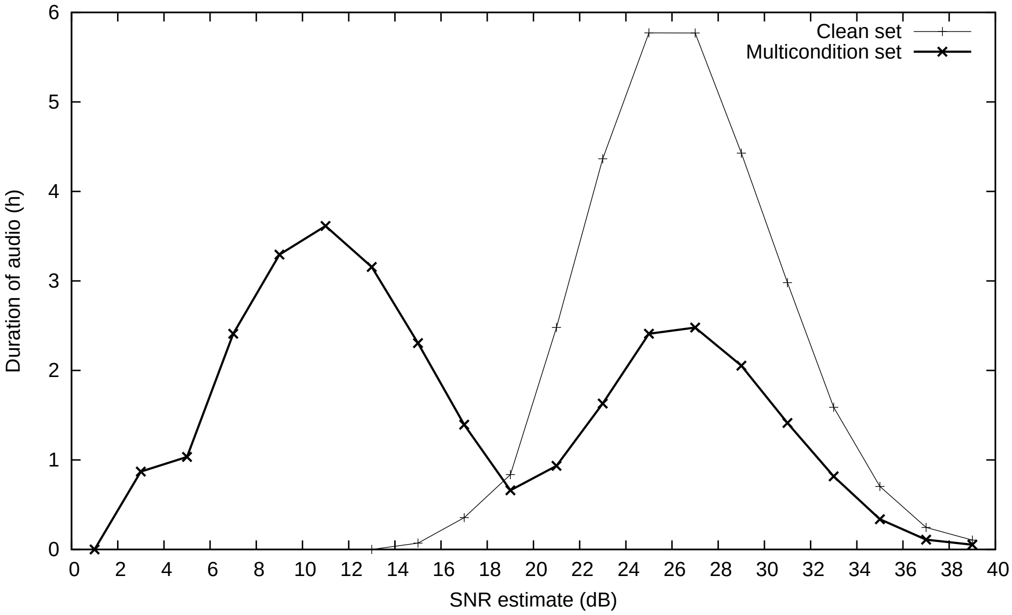

Figure 4.1: *SNR estimate distributions for the clean and noisy speech training sets. Each point in the graph shows the total length of all utterances for which the SNR estimate from the SPEECON corpus transcription annotations falls into the 2 dB range centered at that point. The SNR distribution of the noisy training set has two local maxima corresponding to the clean (channel 0) and noisy (channels 1 and 2) speech samples used in the training.*

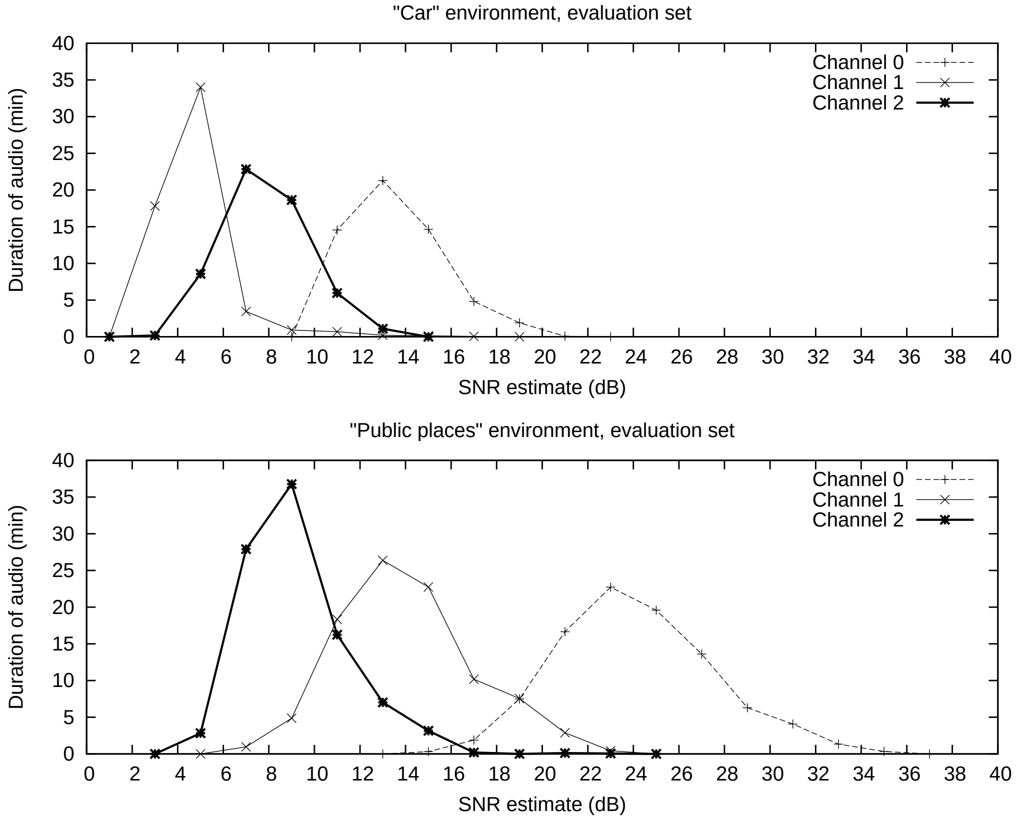

Figure 4.2: *SNR estimate distributions for both "car" and "public places" evaluation sets. Each point in the graph shows the total length of all utterances for which the SNR estimate falls into the 2 dB range centered at that point. SNR estimated similarly to Figure 4.1.*

about the details of this particular recognizer.

The feature extraction stage of the recognizer is based on the MFCC method described in detail in Section 3.2. Unless otherwise noted, a pre-emphasis filter of the form $1 - 0.97z^{-1}$ and frame lengths of 16 ms, with 8 ms overlap between frames, have been used in the experiments. Logarithmic frame energy and the 12 first cepstral coefficients were used as features, with cepstral mean subtraction applied using a window of 150 neighboring frames. First and second derivatives of the MFCC features were computed to yield 39-dimensional feature vectors. Finally, as a post-processing step the features were normalized to have zero mean and unit variance, and a maximum likelihood linear transform (MLLT) estimated during training was applied.

Acoustic modeling of the selected speech features were based on cross-word triphones modeled with state-clustered hidden Markov models (HMM) using Gaussian mixtures. Individual HMM states used a mixture of on average 30 Gaussian distributions to model the speech feature space, and an additional Gamma probability distribution to model the state duration [37]. The language model used by the speech recognizer was a variable-length n-gram model, trained using a growing method [43] on the Finnish Language Bank [1] data set containing book and newspaper data, to a total of approximately 145 million words. The units used in language modeling were statistical morpheme-like units learned with the unsupervised Morfessor method [7] from the text data. Finally, the recognition hypothesis was built with a decoder employing a single-pass time-synchronous Viterbi beam search algorithm [36].

The different feature extraction systems of Chapter 3 have mostly been implemented with MATLAB code, with the exception of the conventional and weighted linear prediction approaches, which were implemented with C++ code directly into the speech recognizer. The main advantage of embedding the SWLP feature extraction directly as a part of the speech recognition was that it removed the cumbersome step of pre-processing the large training audio data into stored SWLP feature vectors, which required a large amount of storage space. In addition, having the SWLP feature extraction implemented in the speech recognizer was seen as potentially useful for later experiments involving the M parameter adaptation. The SWLP feature extraction was naturally more computationally intensive compared to the use of pre-processed features, but the time spent in feature extraction was dominated by the actual speech recognition step.

4.3 Performance evaluation metrics

The speech recognition performance of the different feature extraction methods was measured using letter (LER) and word error rates (WER) of the recognition results. When optimizing parameters of the ASR system in development set experiments, letter error rate was used as the primary performance measure for selecting the best performing system for the final evaluation. The letter error rate was also used for evaluating the differences between systems. Despite being the more common measure for several other languages, the word error rate is not as well suited for Finnish, as Finnish words are often concatenations of several morphemes and correspond to more than one English word. As an example, the Finnish word "kahvin+juoja+lle+kin" translates to "also for a coffee drinker."

The error rate calculation is based on the edit distance between the reference transcript and the recognition hypothesis [30]. The edit distance between two strings is defined as the minimum number of substitution (changing one unit), insertion (adding a unit) and deletion (removing a unit) operations required to transform one of the strings to the other. Denoting the substitution, insertion and deletion error counts by S, I and D, and the total count of units in the reference text by N, the error rate is calculated as

$$\text{error rate} = \frac{S + I + D}{N}. \tag{4.1}$$

For letter and word error rates naturally letters and words, respectively, are used as the units in the edit distance calculation. As the number of insertions is unbounded, the error rates can in exceptional cases exceed 100 %.

The statistical significance of differences between systems was evaluated by performing the Wilcoxon signed-rank test [46] on the letter error rates of the compared pair of systems for corresponding speakers in the evaluation sets. For the test, the individual utterances of the test set are ranked according to the absolute value of the difference between the letter error rates obtained by the systems being compared. The utterance with the smallest absolute value of the difference is assigned rank 1, with successive ranks given to each larger difference, and the ranks are labeled as positive or negative depending on the sign of the difference. The test measure is the smaller value of either the sum of the positive ranks or the sum of the negative ranks. The significance level of $p = 0.05$ was used to classify differences as statistically significant.

4.4 Small-scale experiments

The initial experiments were small-scale tests, using a range of different feature representations prototyped in MATLAB code. In addition to the baseline MFCC features, representations based on linear predictive methods were used. These were based on conventional linear prediction (LP), perceptual linear prediction (PLP) and stabilized weighted linear prediction (SWLP). The relative speech recognition performance of an alternative spectral modeling tool, the minimum variance distortionless response (MVDR) analysis, was also investigated, as well as perceptually motivated ways of utilizing the MVDR estimation.

Most of the evaluated feature extraction methods were based on MFCC features. In LP-MFCC, SWLP-MFCC and MVDR-MFCC feature extraction, the FFT-based magnitude spectrum estimate of traditional MFCC computation was simply replaced with an envelope estimate obtained from an unweighted linear predictive model, a stabilized weighted LP model and a MVDR model, respectively. The perceptual linear prediction (PLP) feature extraction method was not directly based on conventional MFCC features, using instead a more detailed and recent model of the human hearing, as described in Section 3.4. Finally, the PMVDR method is substantially different, as unlike the other methods it does not use a perceptually motivated filterbank approach at all. A more detailed description of all the feature extraction methods can be found in Chapter 3. The PMVDR features were also extracted using tools from the SONIC speech recognizer [34], so different parameter values were used for the pre-processing steps, such as pre-emphasis and framing, as well as the cepstral mean subtraction post-processing step.

The first experiments used a clean speech training set of substantially reduced size, only 111 minutes (1000 utterances), compared to the full training set of over 21 hours (11575 utterances). Also for the recognition, a small development set of a single type of noisy speech from the “car” environment was used. The small data sets were used to make the training and recognition tasks less computationally intensive, as the primary aim was initial development and tuning of the algorithms and their parameters. To conclude the small-scale experiments, a single run with the the full training set was done using the optimal selected parameter values for the different feature extraction methods.

Resulting letter error rates for both the reduced and full training set experiments can be found in Figure 4.3. Results for the unweighted LP-MFCC

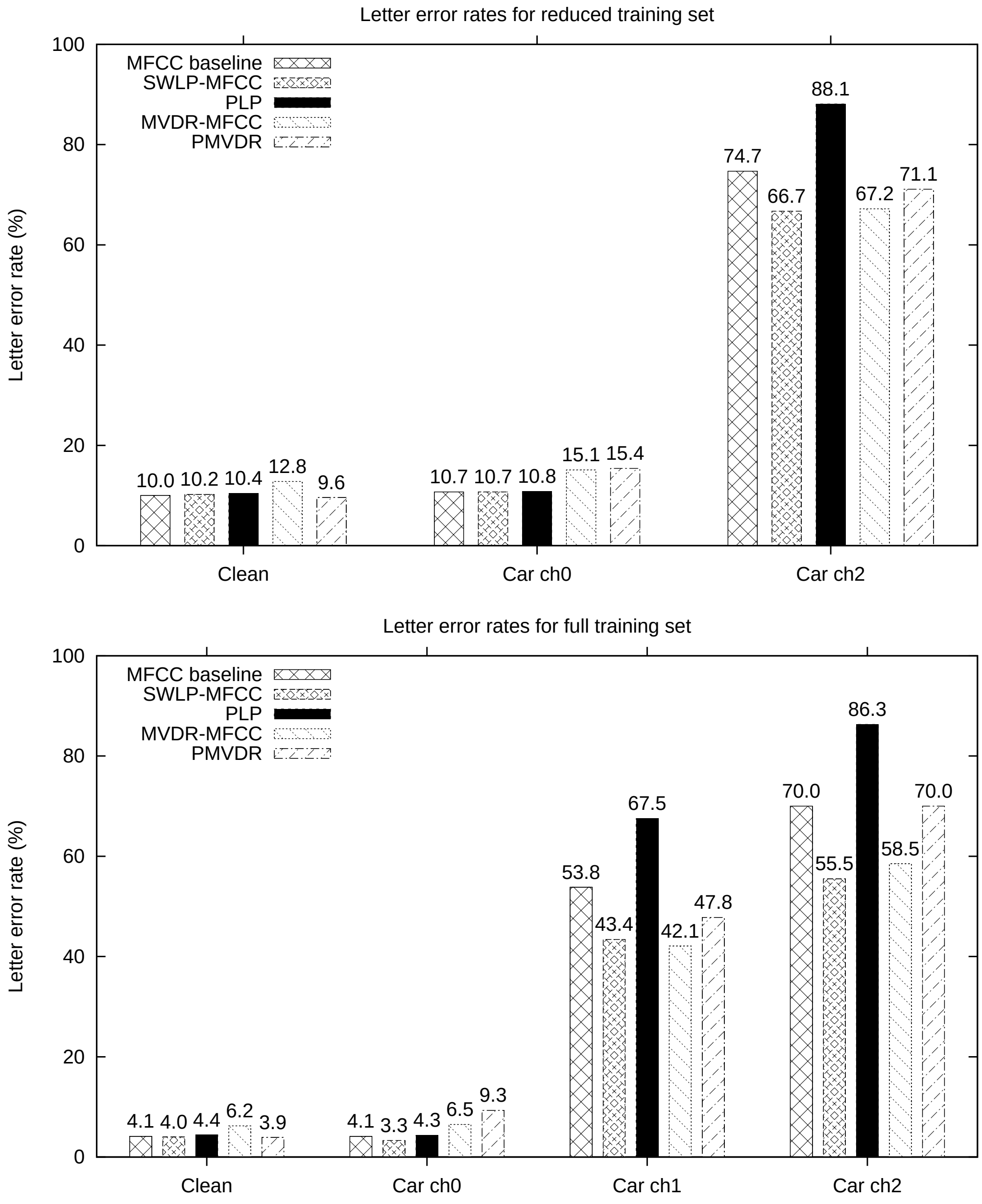

Figure 4.3: *Letter error rates for the initial small scale experiments*

are not included in the figure, as the method was not tested with the full training set at all.

In these tests, the best performing feature extraction system in most cases was the SWLP-MFCC method. The MVDR-MFCC features achieved a similar decrease in letter error rates for the noisier channels 1 and 2, compared to the baseline MFCC model, but incurred some degradation of results when recognizing the clean speech or channel 0. Results for the PLP and PMVDR features were unexpectedly poor.

4.5 Stabilized weighted linear prediction experiments

In these experiments, the main focus was on the speech recognition performance of the SWLP signal modeling method, especially when compared to conventional linear prediction. Some of the preliminary development set experiments were performed also for the MVDR-based feature extraction systems, but due to their unexpectedly low performance, these systems were excluded from the final comparisons. This decision is discussed in Section 5.1.

Initial results and the experiment setup for these experiments were previously presented in a special assignment report by the author of this thesis [22]. The results given in this section are from a later, final test run, where algorithm parameters were optimized with a more comprehensive set of tests with the development set data. These results were also published in our conference paper in the SPECOM 2009 conference [23].

The behavior of the SWLP algorithm depends on the values of the STE weight function window width parameter M, with the use of a short window causing a more prominent emphasis on the energetic regions of the speech frame, and therefore leading to a smoother spectrum estimate. The effect of the M parameter is described in more detail in Section 3.5. In this set of experiments, fixed M parameter values derived from development set experiments were used. Notably, it was found advantageous to use a larger window width in recognition than what was used in training, when recognizing noisier speech with models trained in clean speech only. Preliminary experiences with automatic adaptation of the M parameter are described in sections 4.7 and 4.8.

Full recognition results for the public place and car evaluation sets can be found in Table 4.1. These results compare the performance of the SWLP features (denoted SWLP-MFCC) to conventional linear prediction (LP-MFCC) and the baseline (MFCC) method. Wilcoxon signed rank test was used in pairwise comparisons to find out whether the differences between systems are statistically significant. For significance testing, the "public place" and "car" test sets were combined to a single, larger set. Results of the significance test for the different recording channels and systems, using models trained with clean speech only, can be found in Table 4.2.

From the results it is clear that when using models trained on clean speech to recognize speech recorded in noisy environments, both linear predictive methods show clear improvements in the letter error rate. For the noisy recording channels 1 and 2, these improvements are statistically significant. Furthermore, the recognition results of the SWLP-MFCC system are slightly better than the unweighted LP-MFCC system, reaching statistical significance for the noisier channel 2.

In the multicondition training case, differences between feature extraction systems diminished, with the MFCC baseline system being most often the best performing one. The multicondition training set contained noisy data from both the "car" and "public places" environments described in Section 4.1. While the recordings were independent with respect to the individual noise signals, there was no large mismatch between noise types (e.g. car noise, babble) in the training and test data. However, the noise types used in the training data correspond to very common real-world noisy scenarios, and the test is therefore still quite realistic, compared, for example, to the use of the same artificially added noise sample in both the training and test data, as is done in the popular Aurora noisy speech corpus [16].

4.6 Weighted linear prediction experiments

As described in Section 3.5, the stabilization step performed in the SWLP method has a significant effect on the spectral envelope estimates generated by the weighted linear predictive modeling. A method for extracting a spectral envelope estimate even for a potentially unstable model generated by weighted linear prediction was also presented in Section 3.5. The set of speech recognition experiments described in Section 4.5 were therefore performed also with the unstabilized WLP, in order to investigate if the

Table 4.1: *Letter error rate (word error rate) percentages for the MFCC, LP-MFCC and SWLP-MFCC systems*

"Car" test set, models trained with clean speech.

	channel		
method	0	1	2
MFCC	4.0 *(14.2)*	29.6 *(51.9)*	68.6 *(84.7)*
LP-MFCC	**3.9** *(14.4)*	27.2 *(49.7)*	55.0 *(78.9)*
SWLP-MFCC	4.0 *(14.6)*	**27.1** *(49.5)*	**53.4** *(77.4)*

"Car" test set, models trained with noisy speech.

	channel		
method	0	1	2
MFCC	**3.7** *(14.0)*	**6.8** *(22.1)*	18.0 *(38.3)*
LP-MFCC	3.9 *(14.8)*	7.2 *(23.4)*	**17.6** *(40.1)*
SWLP-MFCC	4.1 *(15.1)*	7.9 *(24.2)*	18.2 *(39.8)*

"Public places" test set, models trained with clean speech.

	channel		
method	0	1	2
MFCC	**3.3** *(13.6)*	23.4 *(41.8)*	40.8 *(56.5)*
LP-MFCC	3.4 *(14.2)*	20.8 *(40.4)*	34.9 *(53.2)*
SWLP-MFCC	**3.3** *(13.6)*	**20.4** *(41.2)*	**33.2** *(53.6)*

"Public places" test set, models trained with noisy speech.

	channel		
method	0	1	2
MFCC	**3.4** *(14.1)*	**6.3** *(21.1)*	**11.9** *(28.8)*
LP-MFCC	3.6 *(14.8)*	7.1 *(23.4)*	12.2 *(30.2)*
SWLP-MFCC	3.7 *(15.0)*	6.7 *(22.1)*	12.0 *(30.0)*

Table 4.2: *Statistical significance results for pairwise comparisons between models using the combined "car" and "public places" data set, with models trained on clean speech only. The name of the better system is shown with the significance level. No statistically significant differences were found for channel 0 results.*

Channel 1

	LP-MFCC	SWLP-MFCC
MFCC	LP-MFCC (p<0.001)	SWLP-MFCC (p<0.05)
LP-MFCC		(p=N.S.)

Channel 2

	LP-MFCC	SWLP-MFCC
MFCC	LP-MFCC (p<0.001)	SWLP-MFCC (p<0.001)
LP-MFCC		SWLP-MFCC (p<0.05)

modification of the spectrum estimate caused by the SWLP stabilization step results in degraded speech recognition performance.

In these experiments, a fixed STE window width value of $M = 16$ was used consistently for all weighted linear prediction tests, as time for the computationally intensive step of selecting the M value based on the results of a large number of development set experiments was not available. The value $M = 16$ was selected based on earlier speech recognition experiments made with WLP in the Department of Signal Processing and Acoustics of Helsinki University of Technology. As can be seen by comparing the results in Table 4.3 with the results presented in Section 4.5, Table 4.1, slightly better performance for the SWLP method can be achieved, based on careful tuning of the M parameter using the development set data. Even the improved SWLP results do not outperform the unstabilized WLP, however.

As the SWLP method was shown to improve results only for the clean speech training case in the experiments of Section 4.5, these experiments omitted the multicondition training tests. Letter error rates for the various systems and recording channels are given in Table 4.3. The MFCC and LP-MFCC baseline results are identical to the results given in Section 4.5. For statistical significance testing, the “car” and “public places” test sets were again combined, as in the experiments of Section 4.5. Results of the statistical significance tests are presented in Table 4.4.

The general trends of the experiment results follow those established in Section 4.5. The linear predictive features outperform the baseline MFCC system for noisy speech, and the difference is more marked for the channel 2 recordings with lower SNR values. The WLP-MFCC method also consistently outperforms the other linear predictive methods LP-MFCC and SWLP-MFCC.

4.7 Oracle-based M adaptation

Since the optimal performance of the SWLP feature extraction method is achieved with a different M parameter value under different noise conditions, automatic adaptation of M is clearly of interest. Theoretical limits for the possible increase in recognition performance that could be obtained by adaptive M selection can be derived using knowledge of the correct recognition result. Considering adaptation where there is a fixed but independent M

Table 4.3: *Letter error rate (word error rate) percentages for the MFCC, LP-MFCC, WLP-MFCC and SWLP-MFCC systems. Only the case of noisy speech being recognized with models trained on clean speech is shown.*

"Car" test set, models trained with clean speech.

	channel		
method	0	1	2
MFCC	4.0 *(14.2)*	29.6 *(51.9)*	68.6 *(84.7)*
LP-MFCC	**3.9** *(14.4)*	27.2 *(49.7)*	55.0 *(78.9)*
WLP-MFCC	4.7 *(16.3)*	**23.1** *(45.5)*	**51.2** *(77.7)*
SWLP-MFCC	4.2 *(15.2)*	32.3 *(54.4)*	55.9 *(77.3)*

"Public places" test set, models trained with clean speech.

	channel		
method	0	1	2
MFCC	**3.3** *(13.6)*	23.4 *(41.8)*	40.8 *(56.5)*
LP-MFCC	3.4 *(14.2)*	20.8 *(40.4)*	34.9 *(53.2)*
WLP-MFCC	4.7 *(18.3)*	**20.0** *(43.3)*	**31.9** *(56.0)*
SWLP-MFCC	3.6 *(14.4)*	20.4 *(41.2)*	34.3 *(53.2)*

Table 4.4: *Statistical significance results for pairwise comparisons between models using the combined "car" and "public places" data set, with models trained on clean speech only. The name of the better system is shown with the significance level. Results are again shown only for the noisier channels 1 and 2.*

Channel 1

	LP-MFCC	WLP-MFCC	SWLP-MFCC
MFCC	LP-MFCC (p<0.001)	WLP-MFCC (p<0.001)	(p=N.S.)
LP-MFCC		WLP-MFCC (p<0.01)	LP-MFCC (p<0.01)
WLP-MFCC			WLP-MFCC (p<0.001)

Channel 2

	LP-MFCC	WLP-MFCC	SWLP-MFCC
MFCC	LP-MFCC (p<0.001)	WLP-MFCC (p<0.001)	SWLP-MFCC (p<0.001)
LP-MFCC		WLP-MFCC (p<0.001)	(p=N.S.)
WLP-MFCC			WLP-MFCC (p<0.001)

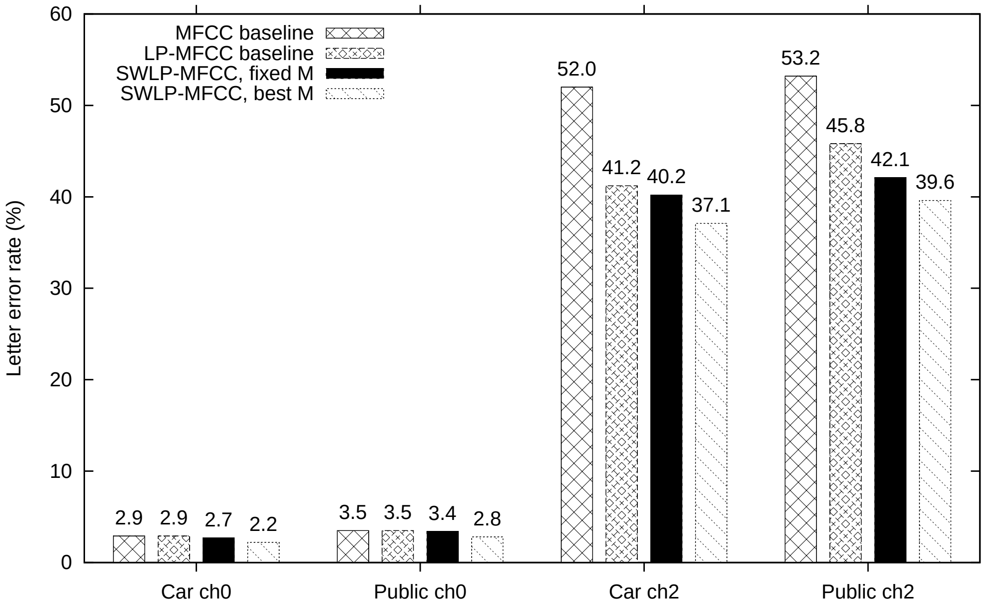

Figure 4.4: *Letter error rates for the oracle-based M adaptation*

parameter value for each recognized utterance, a lower bound for the letter error rate can be estimated by recognizing each utterance with an array of different M values, and computing the letter error rate achieved by selecting the recognition result with least errors. This method is here referred to as oracle-based M adaptation.

The oracle-based M adaptation tests were performed using development set data for the “car” and “public places” environments, channels 0 and 2. Resulting letter error rates can be seen in Figure 4.4. The oracle-based M selection achieves relative letter error rate improvements of approximately 6 % and 18 % over the best fixed M value for recording channels 0 and 2, respectively.

Note that the oracle-based M adaptation was tested in order to seek a theoretical performance bound for adaptive selection of the M value. Our preliminary experiments of automatically selecting an M value which is as close as possible to the oracle-based selection, based on the acoustic model likelihood values, are presented in Section 4.8.

4.8 M adaptation based on acoustic model probabilities

Speech recognition performance of the SWLP features depends strongly on the weight function used. The short-time energy (STE) weights used in this thesis are affected by the window width parameter M, and therefore selecting an M value well-suited to the audio data being recognized is important. The oracle-based M adaptation tests in Section 4.7 suggest that recognition performance can be improved by allowing an independent M value separately for each utterance. For this to be practical, adaptive methods for automatically estimating suitable M values are required.

Probabilities returned by the acoustic and language models for the recognition hypothesis can be used to judge the reliability of the recognition result [21]. In the performed maximum-probability M adaptation experiments, we have again used the results of recognizing the test data with an array of M values to estimate possible speech recognition performance improvements gained by per-sentence M adaptation. In this case, however, the method for selecting the M values does not depend on knowledge of the correct results.

From the acoustic model we can extract for each audio frame an estimate for the probability of the observed speech features, given the HMM state trajectory of the final recognition result. Our experiments showed that selecting the M value for which the probability of non-silent frames is highest leads to improvements in the letter error rate (LER).

Figure 4.5 shows the recognition results for the recording channel 2 of the development set data for the “car” and “public places” environments. In all cases, a fixed M parameter value was used during training, while the recognition was performed with an array of M values. Results reported in the figure include the average letter error rate of the test set for the individual fixed M parameter values, the theoretical lower bound of selecting each recognized sentence with the least errors (the oracle-based M adaptation of Section 4.7) and finally the error rate obtained by selecting the recognition M value using the criterion based on the acoustic model probabilities. For comparison, the MFCC baseline results are also shown.

For both the “car” and “public places” environments, results with models trained with clean speech only are included. In addition, multicondition training results for the “car” environment are given. The proposed method for selecting the M value is an improvement over the best fixed M value for all three test scenarios, but the improvement in the multicondition training case is not sufficient to surpass the MFCC baseline.

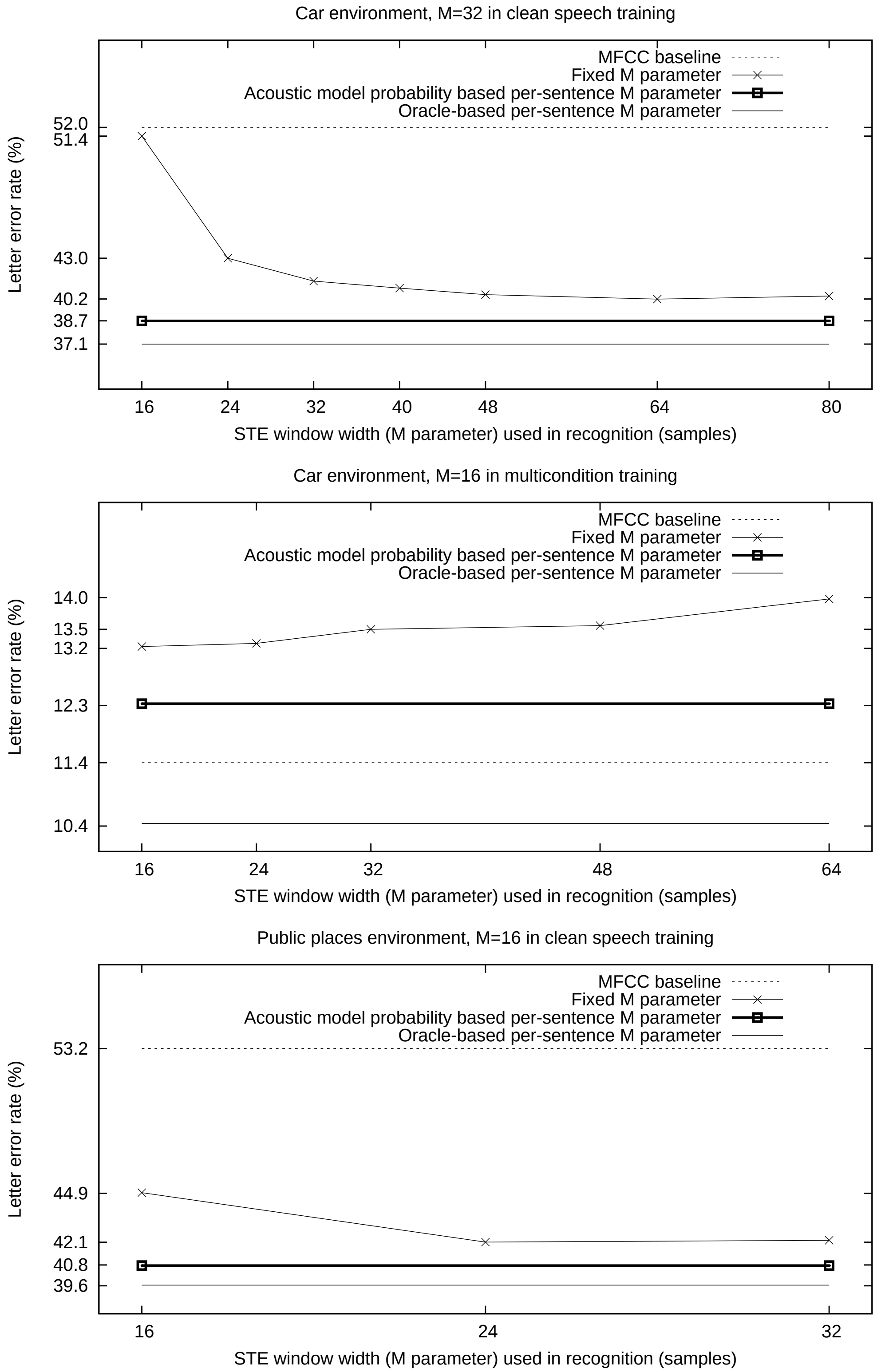

Figure 4.5: *Letter error rates for the acoustic model probability based M adaptation tests on development set data, for both clean speech and multicondition training in the "car" environment, and clean speech training only in the "public places" environment.*

Chapter 5

Discussion

Discussion on the experiments of Chapter 4 is divided to two topics. Section 5.1 considers the spectral envelope estimation method experiments, while Section 5.2 focuses on the parameter adaptation task for the weighted linear prediction approach.

5.1 Spectral envelope estimation models in speech recognition

Results for three sets of experiments comparing the speech recognition performance of feature extraction systems based on different spectrum estimation methods were presented in chapter 4 of this work. Section 4.4 described the results of a preliminary study of a number of dissimilar feature extraction methods. In these experiments, the stabilized weighted linear prediction (SWLP) approach achieved results comparable to feature extraction based on minimum variance distortionless response (MVDR) modeling, which has been studied much in a speech recognition context [10, 31, 48, 49, 50, 51, 52].

After the promising initial results for the SWLP method, a more comprehensive evaluation of its applicability in the feature extraction stage of a large vocabulary continuous speech recognition (LVCSR) system was performed. For these experiments, details such as the filterbank construction for the mel-scale spectrum (see Section 3.2) used by the compared feature extraction systems were made to conform to the MFCC features used by the

Adaptive Informatics Research Centre speech recognizer used in this work. As the change caused an unexpected drop in the performance of the MVDR-based feature extraction methods, they were dropped from later experiments. It was also not possible to alter the parameters used by the Sonic [34] utilities used to generate the PMVDR features. A more thorough comparison of the WLP and MVDR methods remains a possible topic for future work.

In the main set of experiments, the SWLP spectral envelope estimate was compared to two baseline methods. The first baseline method used the FFT periodogram, leading to the common Mel-frequency cepstral coefficient (MFCC) features described in section 3.2, while the second was based on conventional LP. Furthermore, two different sets of training data were tested: one set containing only clean speech, and a multicondition training set including also speech recorded in noisy conditions. Results of this evaluation were presented in section 4.5 and published in [23].

In the case where noisy speech from car and public place environments was recognized using acoustic models trained on clean speech only, both linear predictive methods showed prominent improvement in the recognition results. SWLP was also consistently the better system, with statistically significant improvements for the lower SNR rates. When recognizing clean speech, differences between systems were insignificant, but it should be noted that the use of the SWLP envelope did not cause any degradation of results in this case, which is a common problem with noise robust systems.

Using a multicondition training set containing noisy audio to train the recognition system achieved significant improvements in the recognition performance, a result which agrees well with existing literature [16]. Under these conditions the MFCC baseline generally outperformed the linear predictive methods. It is good to keep in mind, however, than in the section 4.5 experiments the multicondition training set contained noise types from both the car and public place noisy environments used for testing, and while the training and test data were from different recording sessions and speakers, there was no large mismatch between noise types in the data. A recognition experiment with noisy test data of different type could be performed to find out if the robustness of the spectral envelope models could lead to improved recognition performance in that case.

Results from the third set of experiments were presented in section 4.6 and published in [35]. Here the focus was on determining the necessity of the stabilization step of the SWLP method in a speech recognition task, as well as any possible degradation of the results caused by the stabilization. In these experiments, unstabilized weighted linear prediction with the dynamic range limiting detailed in section 3.5 was compared against stabilized weighted linear prediction and the two baseline methods of section 4.5, for the clean speech training scenario. The unstabilized weighted linear was found out to be the best performing system for noisy test material, though at the cost of some degradation of recognition results for clean speech.

The research presented in this work related to weighted linear prediction has focused mostly on the stabilized version, the SWLP algorithm. Based on the results of section 4.6 the unstabilized weighted linear prediction could however be a better fit for speech recognition applications, and is likely to be focused on in further experiments.

Other topics for future work could be based on recent work related to the MVDR spectrum estimation, for example on warped MVDR [48, 49]. The weighted linear prediction approach could be compared against the new MVDR methods. Additionally, ideas from the various refinements of the MVDR method could also be applicable in the weighted linear prediction case.

5.2 Parameter adaptation for the weighted linear prediction

The oracle-based adaptation tests for the STE weight function window width parameter M presented in Section 4.7 suggest that the speech recognition performance of the SWLP method could be improved by adaptively selecting an M value separately for each recognized utterance. In addition, section 4.8 presented one way of selecting the per-utterance M value without utilizing knowledge of the correct recognition result, as was done in the oracle-based adaptation. The results remain preliminary, however, and much future work could be done in this area.

The adaptation tests in section 4.7 and 4.8 were done on an utterance-level granularity, as this made it possible to use the speech recognition system without any modifications to recognize the entire test set separately using a fixed M value. The test results could then be derived by post-processing the per-utterance recognition results to find out the effect of selecting a particular M for a particular utterance. In a real speech recognition system, however, the M value could by dynamically updated as often as desired, possibly for each individual frame if necessary. This could enable the system to better handle noises that are not stationary for the duration of an entire utterance.

In the comparison between stabilized and unstabilized weighted linear prediction in section 4.6, the unstabilized weighted linear prediction had better speech recognition performance. The experiments in that section were carried out using a fixed M parameter value. The performed M adaptation tests could be repeated for the unstabilized WLP method, and in general the WLP method could be used in further adaptation experiments if it is found to have consistently better performance also in these situations. Furthermore, using the unstabilized WLP method caused some degradation in recognition results for clean speech with no background noise. The possibility for reducing or eliminating this degradation with M parameter adaptation could be investigated.

The acoustic model probability criteria used for M value selection in the experiments of section 4.8 was arrived to through exploratory testing. A wide variety of methods have been proposed for deriving confidence values for the recognition results of a speech recognition system [21]. The suitability of these methods for the M selection task could be examined.

The need to recognize the provided input with an array of k different values of M inherent in the M adaptation scheme used in section 4.8, or any refinements of it using a different confidence measure, naturally causes a k-fold increase in the computational complexity of the recognition task. An important research objective would therefore be to find a more practical adaptation method without this limitation.

In [23], the optimal M values were found to be different for the various recording channels, where the major difference between the channels is the relative amount of noise present in the signal. On-line noise estimation methods such as the one presented in [40] could possibly be used in the selection of the M parameter value.

Finally, in this work, a speaker-independent recognizer with no speaker or environment adaptation was used, as the objective was to compare the spectral envelope estimation methods themselves. Adaptation approaches such as the linear transform based maximum likelihood linear regression (MLLR) have been used to obtain impressive improvements in recognition performance [47]. The possibility of using the WLP features in combination with model-level adaptation could be investigated. Research on adaptation methods could also provide new insight on the WLP parameter adaptation task.

Chapter 6

Conclusions

The original motivation for this thesis was to perform a comprehensive evaluation of the stabilized weighted linear prediction (SWLP) spectral envelope estimation method in a large vocabulary automatic speech recognition application, especially when speech is corrupted by the noise in real-world acoustic environments. Feature extraction methods based on the SWLP algorithm as well as a number of other spectral envelope estimation methods were therefore tested in the feature extraction stage of a state of the art large vocabulary continuous speech recognition system. The SPEECON corpus consisting of speech recorded in real-world noisy environments was used to get reliable results of the practical noise-robustness of the compared methods.

Significantly better recognition performance was obtained with the feature extraction methods based on linear prediction (LP) and stabilized weighted linear prediction (SWLP) when compared to a baseline system based on the popular mel-frequency cepstral coefficient (MFCC) feature representation. In addition, the SWLP method was slightly better than unweighted LP, with more pronounced differences in the recognition rates for the noisier test data. Comparisons between speech features based on the stabilized and unstabilized weighted linear prediction also suggest that the unstabilized WLP approach is a better fit for speech recognition purposes, and should be investigated in further research.

The behavior of the weighted linear prediction models depend on the weight function, and in the case of the short-time energy (STE) weight function used in this work, on the STE window width parameter M. Performance improvements were shown to be possible by allowing separate M parameter values for each utterance in the test data, and promising preliminary results were achieved by a method for selecting the per-utterance M value based on the HMM state probabilities provided by the acoustic model of the speech recognizer. The parameter adaptation task also remains a possible topic for later research.

Bibliography

[1] The Language Bank of Finland. http://www.csc.fi/languagebank.

[2] Aubert, X. L. An overview of decoding techniques for large vocabulary continuous speech recognition. *Computer Speech & Language 16*, 1 (January 2002), 89–114.

[3] Baum, L. E., Petrie, T., Soules, G., and Weiss, N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics 41*, 1 (February 1970), 164–171.

[4] Benítez, M. C., Rubio, A., García, P., and de la Torre, A. Different confidence measures for word verification in speech recognition. *Speech Communication 32*, 1-2 (September 2000), 79–94.

[5] Bäckström, T. *Linear Predictive Modelling of Speech – Constraints and Line Spectrum Pair Decomposition.* PhD thesis, Helsinki University of Technology, 5 March 2004.

[6] Chen, S. F., and Goodman, J. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language 13*, 4 (October 1999), 359–393.

[7] Creutz, M., and Lagus, K. Unsupervised discovery of morphemes. In *Proceedings of Morphological and Phonological Learning Workshop of ACL'02* (7-12 July 2002), pp. 21–30.

[8] Davis, S. B., and Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing 28*, 4 (August 1980), 357–366.

[9] Deller, J. R., Hansen, J. H., and Proaxis, J. G. *Discrete-Time Processing of Speech Signals.* IEEE Press, 2000.

[10] Dharanipragada, S., Yapanel, U. H., and Rao, B. D. Robust feature extraction for continuous speech recognition using the MVDR spectrum estimation method. *IEEE Transactions on Audio, Speech, and Language Processing 15*, 1 (January 2007), 224–234.

[11] El-Jaroudi, A., and Makhoul, J. Discrete all-pole modeling. *IEEE Transactions on Signal Processing 39*, 2 (February 1991), 411–423.

[12] Furui, S. Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transactions on Acoustics, Speech and Signal Processing 34*, 1 (February 1986), 52–59.

[13] Gales, M., and Young, S. The application of hidden Markov models in speech recognition. *Foundations and Trends in Signal Processing 1*, 3 (2007), 195–304.

[14] Gold, B., and Morgan, N. *Speech and Audio Signal Processing.* John Wiley and Sons, Inc., New York, 2000.

[15] Hermansky, H. Perceptual linear predictive (PLP) analysis of speech. *Acoustical Society of America Journal 87* (April 1990), 1738–1752.

[16] Hirsch, H.-G., and Pearce, D. The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millennium"* (Paris, France, 18-20 September 2000).

[17] Hirsimäki, T., Creutz, M., Siivola, V., Kurimo, M., Virpioja, S., and Pylkkönen, J. Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech & Language 20*, 4 (October 2006), 515–541.

[18] Hirsimäki, T., and Kurimo, M. Decoder issues in unlimited Finnish speech recognition. In *Proceedings of the 6th Nordic Signal Processing Symposium (Norsig 2004)* (15-17 June 2005), pp. 121–126.

[19] Holmes, J., and Holmes, W. *Speech Synthesis and Recognition*, second ed. Taylor & Francis, London, 2001.

[20] Iskra, D., Grosskopf, B., Marasek, K., van den Huevel, H., Diehl, F., and Kiessling, A. Speecon – speech databases for consumer devices: Database specification and validation. In *LREC* (2002), pp. 329–333.

[21] Jiang, H. Confidence measures for speech recognition: A survey. *Speech Communication 45*, 4 (April 2005), 455–470.

[22] Kallasjoki, H. A comparison of spectrum estimation methods in noise robust LVCSR feature extraction. T-61.5900 Special Assignment, May 19, 2009.

[23] Kallasjoki, H., Palomäki, K., Magi, C., Alku, P., and Kurimo, M. Noise robust LVCSR feature extraction based on stabilized weighted linear prediction. In *13th International Conference on Speech and Computer (SPECOM 2009)* (St. Petersburg, Russia, 21-25 June 2009), pp. 221–225.

[24] Karjalainen, M. Kommunikaatioakustiikka. Report 51, Helsinki University of Technology Laboratory of Acoustics and Audio Signal Processing, 1999.

[25] Levinson, S. E. Continuously variable duration hidden Markov models for automatic speech recognition. *Computer Speech & Language 1*, 1 (March 1986), 29–45.

[26] Ma, C., Kamp, Y., and Willems, L. F. Robust signal selection for linear prediction analysis of voiced speech. *Speech Communication 12*, 1 (March 1993), 69–81.

[27] Magi, C., Bäckström, T., and Alku, P. Stabilised weighted linear prediction – a robust all-pole method for speech processing. In *INTERSPEECH 2007* (27-31 August 2007).

[28] Magi, C., Pohjalainen, J., Bäckström, T., and Alku, P. Stabilised weighted linear prediction. *Speech Communication 51*, 5 (May 2009), 401–411.

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

[30] McCowan, I., Moore, D., Dines, J., Gatica-Perez, D., Flynn, M., Wellner, P., and Bourlard, H. On the use of information retrieval measures for speech recognition evaluation. IDIAP Research Report 04-73, IDIAP Research Institute, 2005.

[31] Murthi, M. N., and Rao, B. D. All-pole modeling of speech based on the minimum variance distortionless response spectrum. *IEEE Transactions on Speech and Audio Processing 8*, 3 (May 2000), 221–239.

[32] Odell, J. J. *The Use of Context in Large Vocabulary Speech Recognition*. PhD thesis, University of Cambridge, 1995.

[33] Oppenheim, A. V., Willsky, A. S., and Nawab, S. H. *Signals & Systems*. Prentice-Hall, 1997.

[34] Pellom, B. Sonic: The University of Colorado continuous speech recognizer. Technical Report TR-CSLR-01, CSLR, University of Colorado at Boulder, 2001.

[35] Pohjalainen, J., Kallasjoki, H., Palomäki, K., Kurimo, M., and Alku, P. Weighted linear prediction for speech analysis in noisy conditions. In *INTERSPEECH 2009* (Brighton, UK, 6-10 September 2009). Accepted for publication.

[36] Pylkkönen, J. An efficient one-pass decoder for Finnish large vocabulary continuous speech recognition. In *2nd Baltic Conference on Human Language Technologies (HLT'2005)* (4-5 April 2005), pp. 167–172.

[37] Pylkkönen, J., and Kurimo, M. Duration modeling techniques for continuous speech recognition. In *INTERSPEECH 2004* (4-8 October 2004), pp. 385–388.

[38] Rabiner, L. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE 77*, 2 (February 1989), 257–286.

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

[40] Rangachari, S., and Loizou, P. C. A noise-estimation algorithm for highly non-stationary environments. *Speech Communication 48*, 2 (February 2006), 220–231.

[41] Röbel, A., and Rodet, X. Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation. In *Proceedings of the International Computer Music Conference (ICMC'02)* (Göteborg, Sweden, 2002), pp. 122–125.

[42] Schaaf, T., and Kemp, T. Confidence measures for spontaneous speech recognition. In *Proceedings of the ICASSP'97* (1997), pp. 875–878.

[43] Siivola, V., Hirsimäki, T., and Virpioja, S. On growing and pruning Kneser-Ney smoothed *N*-gram models. *IEEE Transactions on Audio, Speech, and Language Processing 15*, 5 (July 2007), 1617–1624.

[44] Smith, S. W. *The Scientist and Engineer's Guide to Digital Signal Processing.* California Technical Publishing, 1997.

[45] Viterbi, A. J. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory 13*, 2 (April 1967), 260–269.

[46] Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bulletin 1*, 6 (December 1945), 80–83.

[47] Woodland, P. C. Speaker adaptation for continuous density HMMs: A review. In *Proceedings of ISCA Tutorial and Research Workshop on Adaptation Methods for Speech Recognition* (Sophia Antipolis, France, August 2001), pp. 11–19.

[48] Wölfel, M. Signal adaptive spectral envelope estimation for robust speech recognition. *Speech Communication 51*, 6 (June 2009), 551–561.

[49] Wölfel, M., and McDonough, J. Minimum variance distortionless response spectral estimation - review and refinements. *IEEE Signal Processing Magazine 22*, 5 (September 2005), 117–126.

[50] Yapanel, U. H., and Dharanipragada, S. Perceptual MVDR-based cepstral coefficients (PMCCs) for robust speech recognition. In *Proceedings of the ICASSP '03* (6-10 April 2003), vol. 1, pp. 644–647.

[51] Yapanel, U. H., and Hansen, J. H. L. A new perspective on feature extraction for robust in-vehicle speech recognition. In *INTERSPEECH 2003* (1-4 September 2003), pp. 1281–1284.

[52] Yapanel, U. H., and Hansen, J. H. L. A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition. *Speech Communication 50*, 2 (February 2008), 142–152.