Department of Signal Processing and Acoustics

Robust Methods for Speech Feature Extraction

Jouni Pohjalainen

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 203/2014

Robust Methods for Speech Feature Extraction

Jouni Pohjalainen

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Electrical Engineering, at a public examination held at the lecture hall S1 of the school on 15 December 2014 at 12.

Aalto University
School of Electrical Engineering
Department of Signal Processing and Acoustics

Supervising professor
Professor Paavo Alku

Thesis advisor
Professor Paavo Alku

Preliminary examiners
Professor Kuldip K. Paliwal, Griffith University, Australia
Associate Professor Björn W. Schuller, Imperial College London, UK

Opponent
Professor John H. L. Hansen, The University of Texas at Dallas, USA

Aalto University publication series
DOCTORAL DISSERTATIONS 203/2014

ISBN 978-952-60-6005-7 (printed)
ISBN 978-952-60-6006-4 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-6006-4

Unigrafia Oy
Helsinki 2014

Finland

441 697
Printed matter

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Jouni Pohjalainen

Name of the doctoral dissertation
Robust Methods for Speech Feature Extraction

Publisher School of Electrical Engineering

Unit Department of Signal Processing and Acoustics

Series Aalto University publication series DOCTORAL DISSERTATIONS 203/2014

Field of research Speech and language technology

Manuscript submitted 2 June 2014 **Date of the defence** 15 December 2014

Permission to publish granted (date) 22 September 2014 **Language** English

☐ **Monograph** ☒ **Article dissertation (summary + original articles)**

Abstract

Speech carries information related to, e.g., the linguistic message, speaker identity, speaking situation, speaking style and speaker-related characteristics. Feature extraction refers to the process of converting the digital speech signal into acoustic parameters that can be used to automatically uncover such information, especially using machine learning systems that have been trained on speech data labeled with target information. Such analyses are central in automatic speech recognition, speaker recognition, speech event detection and computational paralinguistic analysis. Each of these application categories is covered in this thesis. With increasing computational and storage capacity of communication technology, speech applications become more widespread and are used in more challenging environments. Ambient noise, varying communication and recording channels and large speaker-related variability tend to cause variation in the acoustic feature statistics and thus mislead speech analysis systems. This study aims to improve the robustness of these systems through feature extraction, so that they better maintain their performance level with increased signal variability.

In short-time feature extraction, the focus is on robust spectrum analysis using especially time-weighted linear predictive methods, in which temporal locations of the signal are differently emphasized. These methods are found to improve additive-noise robustness in automatic speech, speaker and emotion recognition and to improve fundamental-frequency or vocal-effort robustness in formant analysis and speaker recognition. In addition, emphasis of the spectral fine structure is found to improve the robust detection of shouted speech in ambient-noise conditions. In long-term feature processing, modulation filtering of short-time features using multiple time scales is used to emphasize the typical long-term modulation dynamics of a given speech signal class in detecting emotions over a telephone channel in the presence of noise. Feature selection methods capable of tackling data sets with high dimensionality are developed and applied to find relevant utterance-level features to parametrize speech in different paralinguistic tasks with considerable speaker-related variability.

The studies presented have developed speech feature extraction methods that succeed in improving the robustness of various speech analysis systems by focusing on relevant information and de-emphasizing or ignoring irrelevant information. These general-purpose modeling methods are not constrained to any particular application or system structure and thus have many potential uses.

Keywords speech processing, machine learning, robust features, linear prediction

ISBN (printed) 978-952-60-6005-7 **ISBN (pdf)** 978-952-60-6006-4

ISSN-L 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2014

Pages 205 **urn** http://urn.fi/URN:ISBN:978-952-60-6006-4

Aalto-yliopisto

Tiivistelmä

Aalto-yliopisto, PL 11000, 00076 Aalto www.aalto.fi

Tekijä
Jouni Pohjalainen

Väitöskirjan nimi
Robusteja menetelmiä puheen piirrelaskentaan

Julkaisija Sähkötekniikan korkeakoulu

Yksikkö Signaalinkäsittelyn ja akustiikan laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 203/2014

Tutkimusala Puhe- ja kieliteknologia

Käsikirjoituksen pvm 02.06.2014 **Väitöspäivä** 15.12.2014

Julkaisuluvan myöntämispäivä 22.09.2014 **Kieli** Englanti

☐ **Monografia** ☒ **Yhdistelmäväitöskirja (yhteenveto-osa + erillisartikkelit)**

Tiivistelmä

Puhe sisältää informaatiota puhutusta tekstistä, puhujan henkilöllisyydestä, puhetilanteesta, puhetyylistä sekä puhujakohtaisista ominaisuuksista. Piirrelaskennassa digitaalinen puhesignaali muunnetaan akustisiksi parametreiksi, joiden avulla voidaan automaattisesti päätellä mainitun kaltaista informaatiota, erityisesti hyödyntäen koneoppimisjärjestelmiä jotka on opetettu puhemateriaalilla jossa kyseinen informaatio on merkitty. Nämä analyysit ovat keskeisiä automaattisessa puheen- ja puhujantunnistuksessa, puhetapahtumien havainnoinnissa sekä automaattisessa paralingvistisessa analyysissä. Kutakin näistä sovellustyypeistä käsitellään tämän väitöskirjan julkaisuissa. Kommunikaatioteknologian laskenta- ja tallennuskapasiteetin lisääntyessä puheteknologiasovellukset yleistyvät ja niitä käytetään yhä haastavammissa ympäristöissä. Taustamelu, vaihtelevat äänitys- ja siirtokanavat sekä puhujakohtainen vaihtelu aiheuttavat akustisten piirteiden tilastollisten ominaisuuksien vaihtelua ja siten johtavat puheanalyysijärjestelmiä harhaan. Tämän tutkimuksen tavoite on parantaa näiden järjestelmien ns. robustisuutta piirrelaskennan avulla siten, että ne säilyttävät suorituskykynsä paremmin signaaliin liittyvän vaihtelun lisääntyessä. Lyhyen aikavälin piirrelaskennassa keskitytään robusteihin spektrianalyysimenetelmiin käyttäen erityisesti aikapainotettua lineaarista ennustamista, jossa signaalin ajanhetkiä painotetaan eri tavoin. Nämä menetelmät parantavat taustamelurobustisuutta automaattisessa puheen-, puhujan- ja tunnetilojen tunnistuksessa ja perustaajuus- tai puhevoimakkuusrobustisuutta formanttianalyysissa ja puhujantunnistuksessa. Lisäksi spektrin hienorakenteen korostaminen parantaa huudetun puheen havainnointia meluisissa olosuhteissa. Piirteiden modulaatiosuodatus useilla pitkän aikavälin aikaskaaloilla korostaa puhesignaalin luokkien tyypillisiä modulaatiotaajuuksia tunnetilojen havainnoinnissa puhelinpuheessa taustamelun läsnäollessa. Lisäksi tutkitaan piirrevalintamenetelmiä jotka soveltuvat moniulotteisten piirre-esitysmuotojen käsittelyyn. Niitä käytetään etsimään tärkeimmät pitkän aikavälin piirteet paralingvistisissä ongelmissa joissa puhujasta riippuvainen vaihtelu on suurta.

Tässä esitetyissä tutkimuksissa on kehitetty puheen piirrelaskentamenetelmiä, jotka onnistuvat parantamaan erilaisten puheanalyysijärjestelmien robustisuutta keskittymällä oleelliseen informaatioon ja vähentämällä epäolennaisen informaation painoarvoa. Nämä yleisluontoiset mallinnusmenetelmät eivät ole sidottuja mihinkään tiettyyn sovellukseen ja niillä on siten monia mahdollisia käyttökohteita.

Avainsanat puheenkäsittely, koneoppiminen, robustit piirteet, lineaarinen ennustaminen

ISBN (painettu) 978-952-60-6005-7	**ISBN (pdf)** 978-952-60-6006-4	
ISSN-L 1799-4934	**ISSN (painettu)** 1799-4934	**ISSN (pdf)** 1799-4942
Julkaisupaikka Helsinki	**Painopaikka** Helsinki	**Vuosi** 2014
Sivumäärä 205	**urn** http://urn.fi/URN:ISBN:978-952-60-6006-4	

Preface

The present work was carried out at the Department of Signal Processing and Acoustics at Aalto University. It started in 2009 in an Academy of Finland project and was later funded also by the EC FP7 project Simple4All.

I am most grateful to my supervisor, professor Paavo Alku, for his guidance and the opportunity to be working on speech and audio machine learning topics that I find fascinating. His great experience and vision in the field of speech signal processing have been essential for the studies and publications and have allowed me to conduct research on a high international level. The expert knowledge and viewpoints that have been available in both the work environment and through collaborative efforts have been an enormous help in moving forward along this path. I am grateful for the essential contributions of my friend and colleague Carlo Magi (in memoriam, 1980-2008), with whom we began these studies and the ongoing research on feature robustness. I thank all my co-authors and collaborators over the years for their important contributions.

I thank the distinguished pre-examiners Kuldip K. Paliwal and Björn W. Schuller for their valuable comments and feedback on this thesis.

I thank all my current and past colleagues at the department (comprising the earlier Laboratory of Acoustics and Audio Signal Processing), who have contributed to the positive work environment and made working there such a fun experience. I have enjoyed the coffee breaks, the parties and nights out, sports and table football as well as the many conference trips, the holiday trips, paddling and even the occasional research-related discussion. Both research teams I have worked in, as well as the lab in general, have had a great atmosphere and I want to thank everyone involved. During this time, I have made friends within and outside of the lab. Several of these people appear also as co-authors of the papers collected in this thesis. The contributions of my thesis co-authors and their expertise in their specialization areas are much

appreciated.

I thank my friends both near and far away for nice times, visits abroad, hospitality and kindness. I thank the one who, regardless of the current distance, nevertheless has been next to me ever since the early stages of this work. This has been important. We have also spent long hours studying together and she has helped me with many things, including visual details of my conference presentations and the book cover. I thank my family and my relatives for all the support I have received as well as the meetings and get-togethers. For one, I want to mention my grandmother, as I have always been welcome to drop by on my way home for a cup of tea and talk about nice memories from the past as well as current matters. All such things have been important for me and also helpful for maintaining energy in order to proceed with the work. Finally, I am especially grateful to my parents for all their support over the years. You all have greatly helped me to succeed in finishing this thesis.

Espoo, November 20, 2014,

Jouni Pohjalainen

Contents

List of publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Carlo Magi, Jouni Pohjalainen, Tom Bäckström and Paavo Alku. Stabilised weighted linear prediction. *Speech Communication*, vol. 51, no. 5, pp. 401–411, April 2009.

II Rahim Saeidi, Jouni Pohjalainen, Tomi Kinnunen and Paavo Alku. Temporally weighted linear prediction features for tackling additive noise in speaker verification. *IEEE Signal Processing Letters*, vol. 17, no. 6, pp. 599–602, June 2010.

III Jouni Pohjalainen, Tuomo Raitio, Santeri Yrttiaho and Paavo Alku. Detection of shouted speech in noise: human and machine. *Journal of the Acoustical Society of America*, vol. 133, no. 4, pp. 2377–2389, April 2013.

IV Jouni Pohjalainen and Paavo Alku. Extended weighted linear prediction using the autocorrelation snapshot–a robust speech analysis method and its application to recognition of vocal emotions. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*, pp. 1931–1935, Lyon, France, August 25–29, 2013.

V Paavo Alku, Jouni Pohjalainen, Martti Vainio, Anne-Maria Laukkanen and Brad Story. Formant frequency estimation of high-pitched vowels using weighted linear prediction. *Journal of the Acoustical Society of America*, vol. 134, no. 2, pp. 1295–1313, August 2013.

VI Jouni Pohjalainen, Cemal Hanilçi, Tomi Kinnunen and Paavo Alku. Mixture linear prediction in speaker verification under vocal effort mismatch. *IEEE Signal Processing Letters*, vol. 21, no. 12, pp. 1516–1520, December 2014.

VII Jouni Pohjalainen and Paavo Alku. Multi-scale modulation filtering in automatic detection of emotions in telephone speech. In *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014)*, pp. 980–984, Florence, Italy, May 4–9, 2014.

VIII Jouni Pohjalainen, Okko Räsänen and Serdar Kadioglu. Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits. *Computer Speech and Language*, vol. 29, no. 1, pp. 145–171, January 2015.

Author's contribution

Publication I: "Stabilised weighted linear prediction"

The thesis author implemented the mel-frequency cepstral feature extraction based on the stabilized weighted linear prediction (SWLP) spectrum model and the isolated word recognition system based on dynamic time warping. He also implemented the evaluation framework designed to study additive noise mismatch using artificially noise-corrupted test material, analyzed noise statistics and conducted the word recognition experiments.

Publication II: "Temporally weighted linear prediction features for tackling additive noise in speaker verification"

This study was initiated based on initial positive results obtained by the thesis author in limited speaker identification experiments and objective spectral distortion evaluations in the presence of additive noise corruption. The thesis author implemented the mel-frequency cepstral feature extraction based on weighted linear prediction (WLP) and reimplemented SWLP feature extraction using his new formulation of the SWLP method, which appears in this publication. In later publications, including Publication IV, this formulation is seen to form the basis of extended weighted linear prediction (XLP).

Publication III: "Detection of shouted speech in noise: human and machine"

The thesis author discovered the topic, participated in planning and recording the evaluation material, designed and implemented the detection system, carried out the detection experiments and participated in planning the listening

tests. The co-authors contributed significantly in obtaining the test material, implementing the listening tests and analyzing the experimental results.

Publication IV: "Extended weighted linear prediction using the autocorrelation snapshot–a robust speech analysis method and its application to recognition of vocal emotions"

The proposed general formulation was invented jointly by both authors of the paper based on an earlier, less general generalization of weighted linear prediction proposed by the thesis author. The thesis author invented the weighting schemes studied in this paper, implemented the emotion recognition system and conducted the classification experiments.

Publication V: "Formant frequency estimation of high-pitched vowels using weighted linear prediction"

The thesis author contributed to the formulation of the weighted linear predictive methods.

Publication VI: "Mixture linear prediction in speaker verification under vocal effort mismatch"

The thesis author invented and implemented the proposed general method for robust spectrum analysis and designed its present special case designed to tackle formant bias caused by high fundamental frequency.

Publication VII: "Multi-scale modulation filtering in automatic detection of emotions in telephone speech"

The thesis author invented the proposed method, implemented it and conducted the experiments.

Publication VIII: "Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits"

The thesis author initiated the study, invented the unsupervised feature evaluation algorithms and participated in the development of other feature selection methods. The thesis author also invented the methods for combining selection results and for choosing optimal feature subset size and conducted the experiments.

List of abbreviations

AR	autoregressive
ASR	automatic speech recognition
CART	classification and regression trees
CMS	cepstral mean subtraction
CMVN	cepstral mean and variance normalization
CRLP	cepstral residual linear prediction
DCT	discrete cosine transform
DET	detection-error tradeoff (curve)
DNN	deep neural network
ROC	receiver operating characteristics (curve)
DFT	discrete Fourier transform
DTW	dynamic time warping
EER	equal error rate
EM	expectation-maximization
F0	fundamental frequency of speech
FDLP	frequency-domain linear prediction
FFT	fast Fourier transform
GCI	glottal closure instant
GMLP	Gaussian mixture linear prediction
GMM	Gaussian mixture model
HMM	hidden Markov model
IDFT	inverse discrete Fourier transform
IWR	isolated word recognition
KKZ	a cluster initialization method
kNN	k nearest neighbors
LLD	low-level descriptor
LP	linear prediction
LVCSR	large vocabulary continuous speech recognition

MFCC	mel frequency cepstral coefficient
ML	machine learning
MMSE	minimum mean squared error (speech enhancement)
MRMR	minimal-redundancy-maximal-relevance (feature selection)
MVDR	minimum variance distortionless response
NAP	nuisance attribute projection
PDF	probability density function
PLP	perceptual linear prediction
PNCC	power-normalized cepstral coefficients
RASTA	relative spectral (filtering)
RSFS	random subset feature selection
SBE	sequential backward elimination
SD	statistical dependency (feature selection)
SFFS	sequential floating forward selection
SFS	sequential forward selection
SNR	signal-to-noise ratio
SPL	sound pressure level
SSCP	supervised-classification set covering problem
STE	short-time energy
SVM	support vector machine
SWLP	stabilized weighted linear prediction
TMTF	temporal modulation transfer function
UAR	unweighted average recall
UBM	universal background model
VAD	voice activity detection
VTLN	vocal tract length normalization
VQ	vector quantization
WLP	weighted linear prediction
XLP	extended weighted linear prediction
XLP-P	extended weighted linear prediction using partial weights
XLP-S	extended weighted linear prediction using snapshot weights

Abbreviations appearing in the publications only

AME	attenuated main excitation
AVS	absolute value sum
CFS	correlation-based feature selection
CRWLP	cepstral residual weighted linear prediction
DAM	distribution alignment and matching
DAP	discrete all-pole modeling
ILP	integer linear programming
LDA	linear discriminant analysis
LPRA	linear prediction using refined autocorrelation
MI	mutual information (feature selection)
MinDCF	minimum decision cost function
MSLP	Markov-switching linear prediction
PLDA	probabilistic linear discriminant analysis
RBLP	robust linear prediction
SCP	set covering problem
USCP	unsupervised-classification set covering problem

List of symbols

a_k	linear prediction: predictor coefficients
$A(z)$	linear prediction: inverse filter
f	frequency
F_s	sampling frequency
G	linear prediction: gain factor
n	discrete sample or vector index
N	number of observations
p	linear prediction: prediction order
s_n	digital time-domain speech signal
W_n	weighted linear prediction: weighting function
Δ	first-order delta coefficients
$\Delta\Delta$	second-order delta coefficients
λ	parameter set of a mixture model
$\boldsymbol{\mu}_i$	mean vector of the ith mixture component
$\boldsymbol{\Sigma}_i$	covariance matrix of the ith mixture component

List of figures

1. Introduction

The goal of machine learning systems is to infer values of unobserved variables and attributes on the basis of observed variables called *features* [4, 10, 34, 127]. In speech processing, the domain of this thesis, examples include automatic determination of attributes, such as speaker identity or speaking style, based on the digital speech signal. At the more complex end of the problem spectrum, automatic speech recognition is a speech machine learning application that aims to uncover the spoken message.

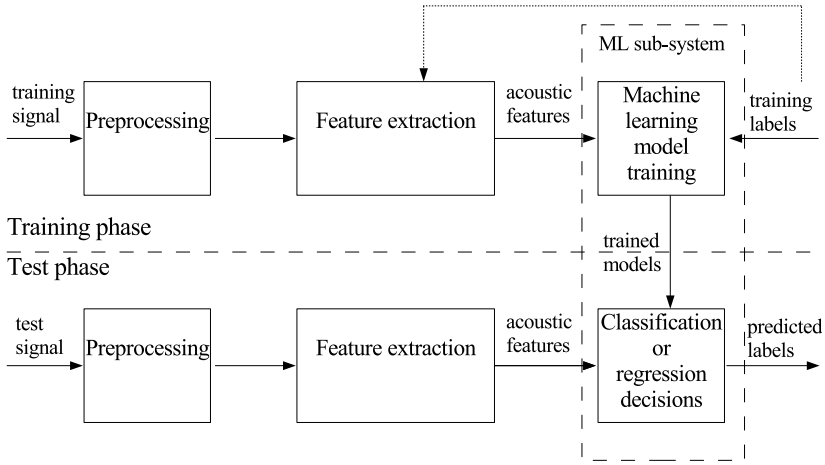

Figure 1.1. Phases of an audio signal analysis system using supervised machine learning (ML).

Figure 1.1 shows the phases of a typical machine learning (ML) system for audio input. A central problem in the design of any ML system is the choice of the feature set: what features should be used to parametrize the observed raw data in order to reliably determine the values of the associated unobserved variables? One might assume that using a high-dimensional, lossless feature

representation, for example the raw data itself, would lead to optimal inference. In practice this is often not the case, because the training data available to train the ML system is frequently too limited both in total amount and in coverage of different conditions. When the dimensionality of the feature space increases, the expected number of data points inside a volume unit decreases. This so-called *curse of dimensionality* [10, 34] means, among other things, that the statistical reliability of a higher-dimensional ML system (i.e., using more features) easily becomes weaker than that of a lower-dimensional ML system, trained with less features but based on the same original raw data. In such a case, the high-dimensional system *overlearns* the training data because of having too many free parameters to account for the relatively small amount of actual data points. On the other hand, a low-dimensional feature representation may not preserve all the information required for optimal performance. Therefore, intelligent feature generation, processing and selection are important practical problems.

This thesis discusses techniques for the generation, transformation and selection of acoustic features in various speech analysis tasks. Its focus is on *robustness* against *mismatch* in ML applications, another effect of limited training data: the limited training material fails to (sufficiently) cover all of the conditions in which the ML system will be used. A common example in speech processing is robustness with respect to acoustic noise: the ML system is trained to process clean (not noisy) speech but applied in a noisy environment. The presence of noise in the test material causes a condition mismatch which is reflected as differences in the feature statistics between the training and test conditions.

The eight studies on speech feature extraction, which comprise this thesis, attempt to alleviate mismatch and degradation conditions related to additive noise, transmission channel and speaker-related aspects. The applications studied are shout event detection in a noisy environment (Publication III), recognition of emotions in a noisy environment (Publication IV) and over the telephone (Publication VII), word recognition (Publication I), text-independent speaker verification (Publications II and VI), formant analysis (Publication V) and automatic voice-based analysis of personality traits and characteristics (Publication VIII). The approaches adopted include spectrum analysis methods which can be focused on the relevant information (Publications I-VI), a feature filtering technique to emphasize class-characteristic modulation frequencies (Publication VII) and novel feature selection techniques to tackle large feature sets (Publication VIII).

Before presenting the main studies, the basics of feature extraction in the domain of digital speech processing, together with their motivations from the perspective of human auditory perception, are reviewed in Section 2. Section 3 lists the main categories of speech applications where ML–and therefore also feature extraction–is central. Section 4 discusses the causes of mismatch in detail and reviews different approaches to improve robustness with different types of mismatch. Section 5 summarizes the studies found in the end of this thesis and Section 6 presents the main conclusions.

2. Speech feature extraction

In this work, *feature extraction* refers primarily to the feature extraction stage of machine learning applications (Figure 1.1), and secondarily also to automatic computation of interpretable low-level features such as estimates of formants (vocal tract resonances). Human auditory perception is generally considered relatively robust and forms the basis of many feature extraction methods. Its basic mechanisms are reviewed in Section 2.1. Popular feature extraction methods, based in varying degrees on the auditory considerations, are discussed in Sections 2.2-2.5.

2.1 Auditory perception

Many feature extraction and post-processing techniques are based on modeling different aspects of auditory perception, whose basics are briefly reviewed. The human auditory system can be divided into two anatomically and functionally distinct regions: the *peripheral* and the *neural* region. The main approaches to modeling auditory perception in practice are the *physiological* approach, where auditory models are derived from explicit physiological and anatomical knowledge, and the *psychoacoustical* approach, which derives *functional* auditory models based on the results of subjective listening tests. Physiological models are mainly limited to modeling the auditory periphery, whose anatomy and physiology are well known. Due to the lack of precise physiological knowledge, psychoacoustics is the principal approach to model the higher-level neural processing. Because psychoacoustical models are typically relatively simple, and computationally less demanding than physiological models, psychoacoustical (functional) models are usually employed also for modeling peripheral processing.

2.1.1 Peripheral auditory processing

Auditory processing begins in the peripheral region consisting of the *outer*, *middle* and *inner ear*. Figure 2.1 shows a schematic diagram focusing on functional importance.

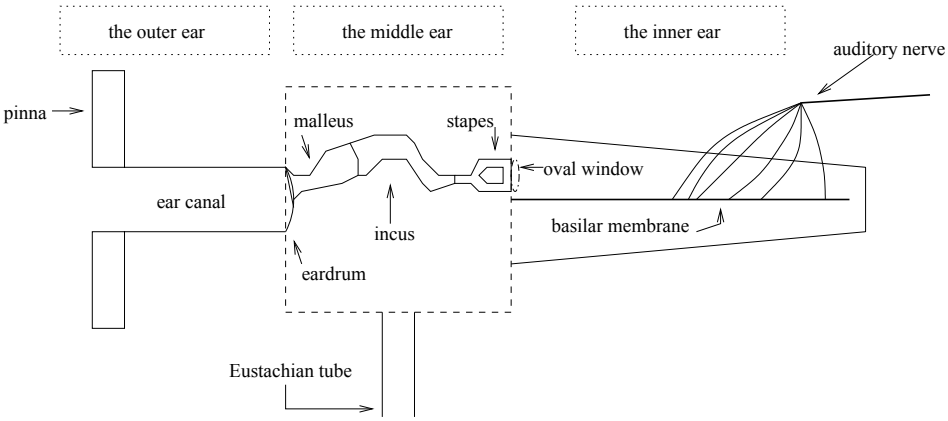

Figure 2.1. Simplified diagram of the human ear (after [67]). Note that the sizes of the different parts are not in proportion and that the part representing the cochlea has been drawn straight instead of coiled.

The outer ear consists of the *pinna*, relevant especially for sound localization on the front-back axis, and the *ear canal*, whose length and width in an adult are roughly 2–3 cm and 0.7 cm, respectively [92]. The ear canal, ending at the *eardrum*, acts as a quarter-wavelength resonator and amplifies energy in the frequency range around 4 kHz [143]. The eardrum marks the beginning of the middle ear, an air-filled cavity of about 6 cm^3 in volume, containing the *ossicular bones*: *malleus* (hammer), *incus* (anvil), and *stapes* (stirrup). Their function in hearing is to linearly transmit eardrum vibrations to the *oval window* membrane of the inner ear. The liquid medium in the inner ear has about 4000 times higher an acoustic impedance than the air medium in the outer ear, and the ossicular bones also perform acoustic impedance transformation. This transformation is based primarily on the large area difference between the eardrum (about 65 mm^2) and the stapes (about 3 mm^2 [67]), but secondarily also on the lever action of the ossicular bones. The impedance match is nearly perfect at frequencies around 1 kHz [143]. The middle ear filter thus contributes, together with ear canal resonance and shadowing and reflection due to the head and the shoulders, to make human hearing particularly sensitive in the approximate frequency range between 1 kHz and 5 kHz. These mechanical filtering effects are reflected in two prominent frequency-dependent psychoacoustical phenomena: the absolute threshold of hearing in quiet and

the *equal loudness* sensation [143]. The absolute threshold of hearing refers to the lowest sound pressure level (SPL) at which a tone is audible at each frequency. The equal loudness contours give, for each frequency of a tone and for the *loudness level* associated with each contour, the SPL at which the tone should be heard in order to sound equally loud as a 1000 Hz tone heard at the SPL equal to the nominal loudness level. Both the SPL and the loudness level are typically expressed in decibels (dB) due to the large dynamic range of hearing.

In the inner ear, of primary importance is the *cochlea*, a coiled tube forming two and a half turns [143]. Filled with two different fluids, it consists of three channels (scalae) that run in parallel from the base to the apex. Sound waves arrive in the cochlea from the middle ear via the oval window, causing vibrations of the fluid called *perilymph* located in the *scala vestibuli*. The vibrations are transferred through the very thin and light *Reissner's membrane* to another fluid called *endolymph* located in the *scala media*. The vibratory motion is transmitted through the endolymph to the *basilar membrane*, along which it proceeds as traveling waves of vertical displacement of the membrane. A traveling wave begins with a small amplitude near the oval window, grows slowly and reaches its maximum at a certain location, after which it rapidly dies out towards the apex of the cochlea. The basilar membrane gradually increases in width and decreases in density along its length of approximately 32 mm [143]. Each location along the basilar membrane responds to sounds of specific frequencies. High-frequency traveling waves resonate near the beginning of the basilar membrane, where it is narrow and stiff, while low frequencies travel across the basilar membrane to resonate near the apex, where the membrane is massive and compliant [92]. The inner ear thus performs frequency separation, or *spectrum analysis*, based on resonance location on the basilar membrane. The resolution of this spectrum analysis is frequency-dependent, being highest at low frequencies. This phenomenon is modeled by different psychoacoustical scales, such as the *Bark* (*critical band*), *ERB* and *mel* scales [67].

The *organ of Corti*, located on the basilar membrane, contains about 30 000 sensory *hair cells* arranged in several rows along the length of the cochlea and the basilar membrane. Basilar membrane vibration in any location causes the affected hair cells to send electrical impulses up the neural fibers of the *auditory nerve*. By measuring these electrical signals, it has been found that the voltage spikes corresponding to stimulation of the hair cells are closely correlated with the mechanical vibration pattern on the basilar membrane up

to frequencies of about 4 or 5 kHz [114]. The *spreading* effect, which manifests itself in this *excitation pattern* of the basilar membrane and can be modeled using functional approximations [126], gives rise to the psychoacoustical phenomenon of *frequency masking*. The instantaneous basilar membrane excitation patterns are also closely connected to the perception of loudness [143], which is additionally affected by *temporal integration*, meaning that the loudness perception generally reaches its maximum after approximately 200 ms from the onset of the sound stimulus [66, 143].

The psychoacoustical approach to auditory modeling involves constructing partial models of the functionality of the auditory system. One fundamental concept is that of the critical band, related to the frequency resolution of hearing discussed above. A critical band defines a frequency range for which perception (as measured by psychoacoustical experiments) abruptly changes as a narrowband sound stimulus is modified to have frequency components beyond the band. When two competing sound signals contribute to the energy passing through a critical-band filter, the sound with the higher energy within the critical band dominates the perception and *masks* the other sound. Critical bandwidths can be measured by various, slightly different psychoacoustical tests [67, 92]. Below 500 Hz, the critical bandwidth stays roughly constant at about 100 Hz. For higher frequencies, it increases with the center frequency. The increase is roughly logarithmic above 1 kHz. The critical bands reach bandwidths of 700 Hz when the center frequency is close to 4 kHz. One functional expression for mapping the frequency f, in Hz, onto the critical-band rate z, or the *Bark scale*, is [92]

$$z = 13\tan^{-1}(0.76f/1000) + 3.5\tan^{-1}(f/7500)^2. \tag{2.1}$$

The *mel scale* is another, practically important, concept of functional modeling of the auditory frequency resolution. It can be approximated by [67]

$$B(f) = 2595\log_{10}(1 + f/700). \tag{2.2}$$

The inverse of the mel scale is thus given by

$$B^{-1}(b) = 700(10^{b/2595} - 1). \tag{2.3}$$

2.1.2 Neural auditory processing

After the initial auditory spectrum analysis on the basilar membrane, recorded by the hair cells, sound information continues on the neural level ascending

through several *nuclei* until the *auditory cortex*. In general, the nuclei along this ascending pathway generate progressively more sophisticated and longer-term representations (with longer memory) of the auditory sensation while also acting as relay stations for lower-level representations from the earlier stages [86, 87]. There is also feedback from the higher neural stages back to lower-level nuclei and to the cochlea. The specialized signal representations generated along the neural pathway involve, for example, modulation frequency selectivity and binaural information related to spatial hearing.

The exact functionality of the different stages of the neural pathway is less well understood than that of the inner ear. The first stage after the cochlea is the *cochlear nucleus* (CN), whose neurons have different time responses: primary-like, onset, chopper, pauser and buildup [87]. Various different types of abstractions of the original auditory stimulus are generated already in specialized regions of the CN. These are passed on through the *superior olivary complex*, which is the initial site of bilateral representation of the acoustic environment, on to the *inferior colliculus* (IC), which is believed to be specialized in the representation of pitch and in localizing sound sources consisting of complex temporal variations. The cells of the IC display *modulation frequency selectivity*, phase-locking to amplitude modulations of the stimulus. Phase-locking to amplitude modulation also occurs in the *medial geniculate nucleus* (MGN), but with lower temporal resolution, i.e., with lower modulation frequencies being represented. After the MGN, the various selective representations of the auditory stimulus reach the auditory cortex. In addition to its role as a relay station on an auditory pathway which conveys all the information necessary to characterize acoustic events, the MGN is also thought to be involved in a second pathway that allows the auditory cortex to selectively label stimuli with perceptual qualities [87]. It may therefore play an essential role in the perception of the acoustic environment and in selective attention in listening.

Psychoacoustical studies have examined the long-term temporal properties of hearing by utilizing various approaches and concepts: temporal integration [143], temporal (forward and backward) masking [66] and the ability of listeners to detect sinusoidal amplitude modulation [7]. The modulation perception studies typically examine the temporal modulation transfer function (TMTF) which describes the sensitivity of hearing to amplitude modulation as a function of frequency. Generally, listeners have been found to be most sensitive to amplitude modulation at modulation frequencies below roughly 10 Hz [7]. In speech, incidentally, most of the modulation energy is concentrated between 2

and 8 Hz and peaks around 4 Hz [49]. Energy in this range is generally largely affected by phonemic and syllabic variation.

At the psychoacoustical level, the perception of sounds can be divided into four components: *loudness*, *pitch*, *timbre* and *subjective duration* [114, 143]. Clear physical correlates can be found, in the properties of physical sounds, for loudness (sound intensity), pitch (fundamental frequency) and subjective duration (true duration). Timbre, which is frequently important in distinguishing between different classes of sounds, is a collective name for many perceptual aspects for which no simple one-dimensional physical correlate can be found [15, 67, 143]. It can be thought of as describing the auditory spectrum, closely related to the instantaneous basilar membrane excitation pattern. Thus, the closest purely physical and signal-related representation of timbre is the short-time magnitude spectrum. In speech and audio processing applications, the size of a short-time *analysis frame*, during which spectral properties are assumed to stay constant, is usually in the range 10–30 milliseconds, most commonly 20–25 milliseconds. The *frame shift interval*, i.e., the temporal distance between two successive analysis frames, can vary according to application but is most typically close to 10 milliseconds.

High-level neural auditory processing is manifested in the psychoacoustical sense as formation of temporal *auditory streams* which, according to Bregman [15], are perceptual representations in terms of which the auditory system organizes incoming auditory evidence. Distinct auditory streams tend to present the evolution of distinct environmental sounds across time. Auditory stream formation is evidently well predicted by the perceptual grouping principles of *Gestalt psychology*. According to Gestalt theory, this organization into perceptual *patterns* (*Gestalt* in German) is governed by a competition of "forces of attraction" between sensory elements. According to the German Gestalt psychologists in the early 20th century, it was impossible to perceive sensory elements without their forming an organized whole. They argued that this tendency of pattern formation is an innate tendency of the brain [15]. Gestalt principles are often encountered in describing visual perception. Auditory analogies presented by Bregman include:

- Principle of proximity: proximity of sound elements in time and/or frequency favors their being grouped into the same auditory stream (compare with spatial proximity in vision).

- Principle of similarity: sounds with similar timbre are more likely to be

grouped into the same stream (compare with visual elements having similar texture). While timbre is necessarily a multidimensional quality (e.g., the auditory spectrum), *brightness*, or the approximate balance of high- and low-frequency energy in the spectrum, has been identified as a central dimension of timbral similarity in auditory stream formation.

- Principle of closure: sounds temporarily masked by other sounds tend to be perceived as continuing during the masked segments (compare with partially occluded objects in an image).

- Principle of common fate: different parts of the short-time spectrum that change in a similar way at the same time tend to be perceptually grouped together (compare with visual elements moving together being perceived as a group). The said change in the spectral components can be frequency modulation (common variation in the center frequencies) or amplitude modulation (common variation in the amplitudes of the frequency components).

According to Bregman, the Gestalt grouping principles can be viewed as heuristics that combine their effects, much like voting, to aid the auditory system in decomposing a mixture of sounds into separate perceptual entities corresponding to different real-world *events*. In [15], grouping principles in auditory stream formation are divided into sequential (temporal) and simultaneous (spectral) grouping. Bregman argues that there are two ways of acquiring skills for auditory stream segregation: *primitive segregation*, which is an innate mechanism, and *schema-based segregation*, which is based on learned *schemas* and likely involves learned control of attention. Considering the principle of similarity and the role of the different dimensions of timbre, brightness appears to be important for primitive segregation of auditory streams; Bregman did not find evidence for other acoustic dimensions related to the timbre perception being used at all in primitive segregation [15].

Allen [3] reviews the research on human speech recognition performed by Fletcher et al. [43] and presents certain conclusions that are in line with Bregman's findings. In particular, compared to across-frequency processing (spectral or timbral templates), human recognition of speech is claimed to rely more on across-time processing, with only local coupling across frequency. This statement is reminiscent of auditory stream formation by the principles of proximity, closure and common fate, as well as only the brightness dimension being important in similarity-based primitive segregation. These perceptual

studies thus highlight the importance of long-term modeling of auditory perception. However, due to neural auditory processing being both complex and less well understood than peripheral auditory processing, long-term modeling is less straightforward than short-term modeling of peripheral processing.

2.2 Spectrum analysis

Given the role of the inner ear as a spectrum analyzer, modeling of the short-time magnitude spectrum is fundamental for auditorily motivated speech signal processing (even though, under certain conditions, it is also possible to apply the *phase spectrum* as an alternative representation with similar perceptual importance [93]). This section reviews two of the most central tools for spectrum analysis that find frequent application in feature extraction as well as in speech, audio and signal processing in general: the discrete Fourier transform and linear prediction.

2.2.1 Discrete Fourier transform

Fourier spectrum analysis is a central mathematical tool in signal processing. For a continuous signal $h(t)$, the *continuous Fourier transform* $H(f)$ and its corresponding inverse transform back to the signal domain are given by

$$H(f) = \int_{-\infty}^{\infty} h(t) e^{-2\pi \mathrm{i} f t} \mathrm{d}t \tag{2.4}$$

$$h(t) = \int_{-\infty}^{\infty} H(f) e^{2\pi \mathrm{i} f t} \mathrm{d}f. \tag{2.5}$$

Here, t denotes a coordinate, typically time, and f denotes the number of cycles per unit of t, i.e., frequency. For discretely sampled sequences of length K[1], the *discrete Fourier transform* (DFT) is given by

$$H_k = \sum_{n=0}^{K-1} h_n e^{-2\pi \mathrm{i} n k / K}. \tag{2.6}$$

It is related to the continuous Fourier transform as $H(f = k/(K\Delta t)) \approx H_k \Delta t$, where Δt is the sampling interval of the discrete sequence h_n [105]. The inverse DFT (IDFT) for a DFT sequence of length K is given by

[1]For K less than or larger than the original number of samples N, the original signal h_n has been truncated or zero-padded, respectively.

$$h_n = \frac{1}{K} \sum_{k=0}^{K-1} H_k e^{2\pi i n k/K} \tag{2.7}$$

A DFT of length K is called symmetric (even) if $H_k = H_{K-k}$ and antisymmetric (odd) if $H_k = -H_{K-k}$. Similarly, a sequence of length K is symmetric if $h_n = h_{K-n}$ and antisymmetric if $h_n = -h_{K-n}$. Some properties of the DFT (which are similar to those of the continuous Fourier transform) include [91, 105]:

1. The DFT/IDFT of a real sequence is conjugate symmetric.

2. The DFT/IDFT of a symmetric sequence is symmetric.

3. The DFT/IDFT of an antisymmetric sequence is antisymmetric.

4. The DFT/IDFT of a real and symmetric sequence is real and symmetric (follows from properties 1 and 2).

5. The magnitude spectrum of a real sequence is symmetric (follows from property 1).

Fast Fourier transform (FFT) algorithms are computationally efficient algorithms for computing the DFT. Many variants of the FFT exist. Their time complexity is generally $O(K \log K)$, instead of $O(K^2)$ that results when using Eqs. 2.6 and 2.7 directly [91, 105].

The power spectrum (or power spectral density) $P(f)$ of a wide-sense stationary random process is defined as the discrete-time Fourier transform of the autocorrelation sequence [53]. The *periodogram* is a simple estimate of the power spectrum of the random process that produces the observed signal h_n. For normalized frequency f, it is given by [53]

$$\hat{P}(f = k/K) = \frac{1}{K} |H_k|^2, \tag{2.8}$$

where H_k is the K-point DFT of h_n. The periodogram spectrum estimation using the FFT is a very widely used method for short-time spectrum analysis. It is also commonly applied in speech feature extraction.

2.2.2 Linear prediction

A linear predictive model represents a time-domain signal s_n parametrically as a pth-order autoregressive process [82, 84, 109]

$$s_n = \sum_{k=1}^{p} a_k s_{n-k} + G u_n, \tag{2.9}$$

whose system function in the complex frequency domain (z domain) is of the *all-pole* form

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} = \frac{G}{A(z)}, \tag{2.10}$$

i.e., it contains only poles outside the origin of the z domain. In the above equations, the coefficients a_k are known as the *predictor coefficients*, G is a gain term and u_n is an excitation signal, with mean zero, acting as input to the system. $A(z)$ is known as the *inverse filter* or, alternatively, the *prediction error filter*.

Based on Eq. 2.9, a prediction of the value of any signal sample s_n can be obtained as a linear combination of the past p samples, $\hat{s}_n = \sum_{k=1}^{p} a_k s_{n-k}$. Filtering s_n by the inverse filter $A(z)$ gives the prediction error signal (residual) $e_n = s_n - \sum_{k=1}^{p} a_k s_{n-k} = s_n - \hat{s}_n$. In order to solve the predictor coefficients, linear prediction (LP) analysis minimizes the energy of the prediction error

$$E = \sum_n e_n^2 = \sum_n \left(s_n - \sum_{k=1}^{p} a_k s_{n-k} \right)^2 \tag{2.11}$$

within the short-time analysis frame consisting of samples s_n. The partial derivatives of Eq. 2.11 with respect to each predictor coefficient a_j are given by

$$\frac{\partial E}{\partial a_j} = 2 \sum_{k=1}^{p} a_k \sum_n s_{n-k} s_{n-j} - 2 \sum_n s_n s_{n-j}, \qquad 1 \le j \le p. \tag{2.12}$$

Setting each of these derivatives to zero yields the normal equations

$$\sum_{k=1}^{p} a_k \sum_n s_{n-k} s_{n-i} = \sum_n s_n s_{n-i}, \quad 1 \le i \le p, \tag{2.13}$$

which are solved to obtain the model parameters, i.e., the predictor coefficients a_k.

LP can be used to model the broad shape, or upper envelope, of the short-time magnitude spectrum as $|H(z)|$. In order to do this, two complex poles are needed to model one spectral peak, such as a formant, over the signal bandwidth $(0, F_s/2)$ determined by the sampling frequency F_s. Thus, two

poles should be allocated for each expected spectrum envelope peak within the signal band. Speech has, on average, one formant per kHz of the signal band $(0, F_s/2)$ [109]. In addition, a few poles are required to model the excitation source and lip radiation, affecting the spectral tilt. Thus, in order to model the spectrum envelope, the order p is typically chosen as the sampling frequency in kHz plus a small integer [109]. For a signal sampled at 16 kHz, for instance, $p = 20$ would be a typical choice. With such a choice, the signal model has a direct connection to the *source-filter model* of speech production [41]: the filter $H(z)$ depicts the vocal tract *filter*, while the excitation signal u_n plays the role of the voiced (glottal) or unvoiced (fricative) *source*. An all-pole model can also be viewed as a digital-filter representation of a lossless acoustic tube model, a simplified physical model of the vocal tract [84]. This highlights the suitability of all-pole models for speech spectrum modeling.

In Eqs. 2.11–2.13, the range of summation of n has not been specified. A proper choice of n gives rise to the two canonical methods of LP analysis: the *autocorrelation method*, in which the prediction-error energy is minimized over a theoretically infinite interval, but s_n is considered to be zero outside the actual analysis window; and the *covariance method*, in which the prediction-error energy is minimized over the interval $(p, N-1)$ when the analysis frame consists of samples s_n, $n \in (0, N-1)$. The synthesis model $H(z) = 1/A(z)$ given by the LP autocorrelation method is guaranteed to be *stable*, meaning that the roots of the denominator polynomial $A(z)$ all lie inside the unit circle [82]. It is thus more suitable for synthesis applications than the covariance method. Moreover, an efficient solution algorithm called the Levinson-Durbin recursion exists for the autocorrelation method [82, 109]. A third notable method for the LP coefficient solution is the lattice-based *Burg's method* [53]. Finally, *adaptive filtering* can also be used to learn a linear prediction model and to update it sample by sample [53, 54].

Two sets of LP filter coefficients can be compared using metrics such as the Itakura distance [64]. Often, however, the direct form filter coefficients are converted to another form before further processing. In pattern recognition and machine learning applications, there are two main approaches to representing the magnitude spectrum information of the linear predictive all-pole filter. The inverse filter FIR coefficients $A(z)$ can be interpreted as an impulse response, and the FFT can be applied on them directly to obtain the spectrum of the inverse filter. This spectrum can be inverted in the frequency domain–taking care to limit the spectral dynamics in case the all-pole coefficient solution is not stable–to get the spectrum of the synthesis filter $H(z)$.

Another approach is to convert the filter coefficients directly to a *cepstral* representation. Cepstral techniques are discussed in Section 2.3.

2.2.3 Spectral envelope and fine structure

The Fourier transform and LP are generally applied to different types of spectrum modeling. The Fourier transform represents the complete magnitude spectrum, including its fine structure. In contrast, LP is normally applied to model the broad spectral shape or, more specifically, the upper envelope of the magnitude spectrum, which is mostly affected by the formants. This is illustrated in Figure 2.2. In speech coding, for example, LP is used to parametrize the overall spectrum shape.

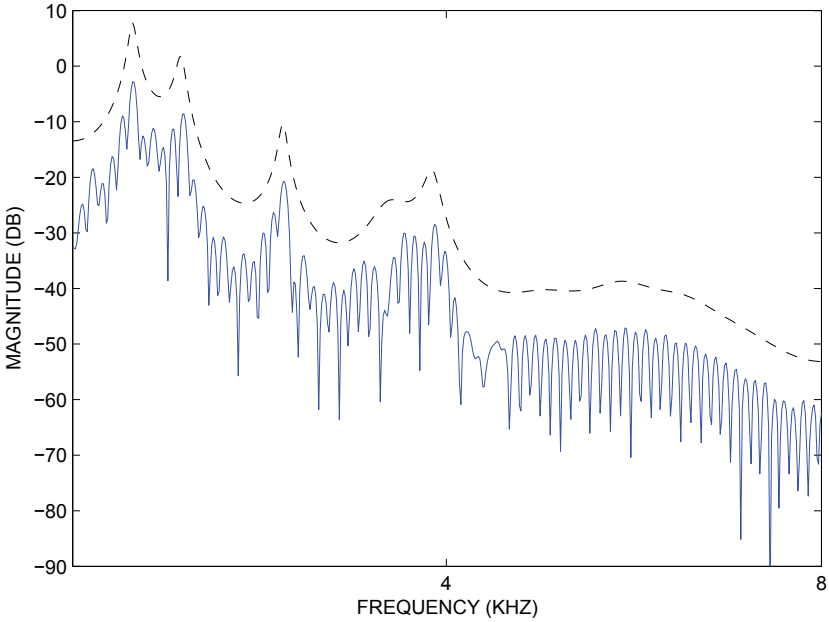

Figure 2.2. Comparison of typical Fourier (solid line) and LP (dashed line) spectra for a vowel frame.

With a high prediction order p, approaching the size N of the (typical) analysis frame, LP could also model the harmonics and the fine structure of the spectrum. However, this would still tend to focus on the spectral peaks rather than valleys (an inherent property of LP, which is due to its goal of minimizing the squared modeling error which, when viewed in the spectral domain, focuses on the high-energy peaks) and would also be computationally less efficient than FFT (FFT is $O(N \log N)$, while the correlation computation in LP is $O(Np)$, and the normal equation solution in the autocorrelation method is $O(p^2)$ [109]).

In many practical applications, modeling the short-term spectrum envelope leads to sufficiently informative features for the speech signal. Exceptions do, however, exist where the fine structure related to the glottal excitation becomes particularly important, such as in Publication III of this thesis.

In addition to linear predictive (all-pole) methods, broad shapes and the fine structure of the spectrum can also be separated using cepstrum analysis, discussed in Section 2.3.

2.3 Cepstrum analysis

Two mutually convolved signal components that have sufficiently different spectra can be decomposed using *cepstrum analysis*, which belongs to the class of *homomorphic* signal processing techniques [91, 109]. For discrete-time sequences, the *real cepstrum* is defined as

$$\{C_j\} = \mathrm{IDFT}\left(\{\log\left(|\mathrm{DFT}(\{s_n\})|\right)\}\right), \tag{2.14}$$

where $\mathrm{DFT}(\{x_n\})$ and $\mathrm{IDFT}(\{x_n\})$ denote the discrete Fourier transform (Eq. 2.6) and the inverse discrete Fourier transform (Eq. 2.7), respectively, performed on some discrete sequence $\{x_n\}$. The input speech signal is denoted by s_n. From the selected properties of Fourier transforms listed in Section 2.2.1, it follows that the real cepstrum is real-valued and symmetric[2].

It is assumed that two time-domain signals, s_n and h_n, have been convolved and the result is observed as one composite signal $r_n = s_n \star h_n$. The goal of the analysis is to find the magnitude spectrum $|S_k|$ of s_n. Applying the convolution theorem [91], the magnitude spectrum of the observed signal is given by

$$|R_k| = |S_k||H_k| \tag{2.15}$$

where $|H_k|$ is the magnitude spectrum of the "channel" h_n. Taking the logarithm (base not relevant) of both sides of the equation gives

$$\log|R_k| = \log|S_k| + \log|H_k|. \tag{2.16}$$

In the logarithmic magnitude spectrum of r_n, an undesired component appears as an additive component $\log|H_k|$ instead of a convolved component. However, the two components still cannot be separated without an estimate of

[2]It is worth pointing out that the name "real cepstrum" does not refer to the fact that it is real-valued. The real cepstrum and the *complex cepstrum* differ in whether the real or complex logarithm is used [91].

$|H_k|$. Suppose, however, that it is known from experience that the logarithmic magnitude spectrum $|H_k|$ of the convolved component will be systematically different from $|S_k|$. Typically, this means that one spectrum changes much more rapidly with frequency than the other. In other words, one of the mutually convolved spectra contains mostly the *fine structure* while the other has a broader shape. Taking the (inverse) DFT of the logarithmic magnitude spectrum yields

$$\{C_j\} = \text{IDFT}\left(\{\log|R_k|\}\right) = \text{IDFT}\left(\{\log|S_k|\}\right) + \text{IDFT}\left(\{\log|H_k|\}\right), \quad (2.17)$$

in which the components due to the two signals, s_n and h_n, remain mutually additive because the IDFT is a linear operation[3]. The two components, originally convolved in the time domain, are now additive and should inhabit different "quefrency" (a term for the cepstral-domain "frequency") ranges. This means they can be easily separated by "liftering", that is, windowing in the cepstral domain (similarly to how conventional filtering is performed in the spectral domain). The low-quefrency components correspond to broad spectrum shapes, while the higher quefrencies contain information about the spectral fine structure.

The root-mean-square (rms) log-spectral distance is known to be a perceptually relevant distance measure between two audio signals because of the inner ear's spectrum analysis and the large dynamic range of human hearing [143]. It can be shown that the Euclidean distance between two cepstra, which have both been liftered using a rectangular low-quefrency window and truncated to the length of the window L, sets a lower bound on the rms log-spectral distance [48]. With the inclusion of more cepstral coefficients by increasing the lifter length L, more spectral fine structure information is included in the truncated cepstrum, and the cepstral Euclidean distance approaches the rms log-spectral distance from below [48]. In audio pattern recognition, the fact that the simple Euclidean distance is perceptually meaningful permits the use of simpler probabilistic models for the feature vectors[4]. Because of this desirable property, the cepstrum is often used as a feature vector representation

[3]The choice of the IDFT instead of the DFT only affects the scaling and not the shape of the real cepstrum. Because the logarithmic magnitude spectrum is real and symmetric, its DFT or IDFT will be real and symmetric; if the DFT/IDFT is real, it is known that the opposite phase between Eqs. 2.6 and 2.7 is canceled out in the summation and does not affect the transform. Thus, the only practical difference between using Eq. 2.6 or 2.7 in computing the real cepstrum is the scaling coefficient $1/K$ in 2.7).

[4]Because the Euclidean distance, which when applied to cepstral vectors approximates the perceptually relevant rms log-spectral distance, is a special case of the

for magnitude spectral models computed by other means, such as LP analysis. Conversion from the filter coefficients of a stable LP filter to cepstral coefficients can be accomplished by the recursive formula [84, 109]

$$c_0 = \log(G) \tag{2.18}$$

$$c_n = a_n + \sum_{m=1}^{n-1} \left(\frac{m}{n}\right) c_m a_{n-m}, \quad 1 \leq n \leq p. \tag{2.19}$$

The conversion of cepstral coefficients to LP coefficients readily follows from the same formulas. It should be noted that the conversion formula will only produce as many coefficients as are present in the original representation. When an LP model, representing the broad spectrum shape, is converted using the formula, the resulting truncated cepstrum will not include high-quefrency cepstral coefficients that would be responsible for parametrizing the spectral fine structure. Cepstral representations often omit the "zeroth" coefficient (Eq. 2.18), which is related to the absolute energy of the signal being modeled.

The first step in the computation of a cepstrum is to obtain a magnitude spectrum estimate, either by a DFT (as in Eq. 2.14) or by LP (conversion formulas given by Eqs. 2.18-2.19). The replacement of such a conventional spectrum model by an *auditory spectrum* leads to a cepstral representation with increased perceptual emphasis, as discussed in Section 2.4.

2.4 Mel-frequency cepstral coefficients

While using logarithmic spectra as features is already perceptually justified, as discussed earlier, this does not yet sufficiently model peripheral auditory processing. The most notable functionality not captured by logarithmic spectra is the non-uniform frequency resolution of hearing. *Auditory filterbanks* implement this *frequency warping*, which occurs on the basilar membrane where sound pressure spectra are nonlinearly mapped to auditory excitation patterns [143]. While sophisticated time-domain filter models have been developed that attempt to duplicate the physiology of the inner ear, simple functional models are typically sufficient for acoustic feature extraction in recognition applications. *Mel-scale filterbanks* are one such approach that has found great success. They are often used to obtain *mel-frequency cepstral coefficients* (MFCCs) [27].

In the frequency domain, a triangular mel-filterbank with M filters, such as the one whose transfer functions are shown in Figure 2.3, can be constructed

Mahalanobis distance, whose minimization is equivalent to maximization of the log likelihood of a multivariate Gaussian distribution [127], a Gaussian assumption for the distributions of cepstral vectors is perceptually justifiable.

as follows. Denote the lowest and highest frequency of the filterbank by f_l and f_h, respectively. Let F_s be the sampling frequency, K the DFT size and L the DFT index corresponding to the Nyquist frequency $F_\mathrm{s}/2$. Determine the frequency boundaries of the filter transfer functions so that they are uniformly spaced on the mel scale [63]:

$$f(m) = \left(\frac{K}{F_\mathrm{s}}\right) B^{-1}\left(B(f_\mathrm{l}) + m\frac{B(f_\mathrm{h}) - B(f_\mathrm{l})}{M+1}\right), \qquad 0 \le m \le M+1, \quad (2.20)$$

where the mel-scale function B and its inverse function B^{-1} are given by Eqs. 2.2 and 2.3, respectively.

The transfer functions of the type shown in Figure 2.3 are given by [63]

$$H_{m,k} = \begin{cases} 0, & k < f(m-1), \\ \frac{(k-f(m-1))}{(f(m)-f(m-1))}, & f(m-1) \le k \le f(m), \\ \frac{(f(m+1)-k)}{(f(m+1)-f(m))}, & f(m) \le k \le f(m+1), \\ 0, & k > f(m+1). \end{cases} \quad (2.21)$$

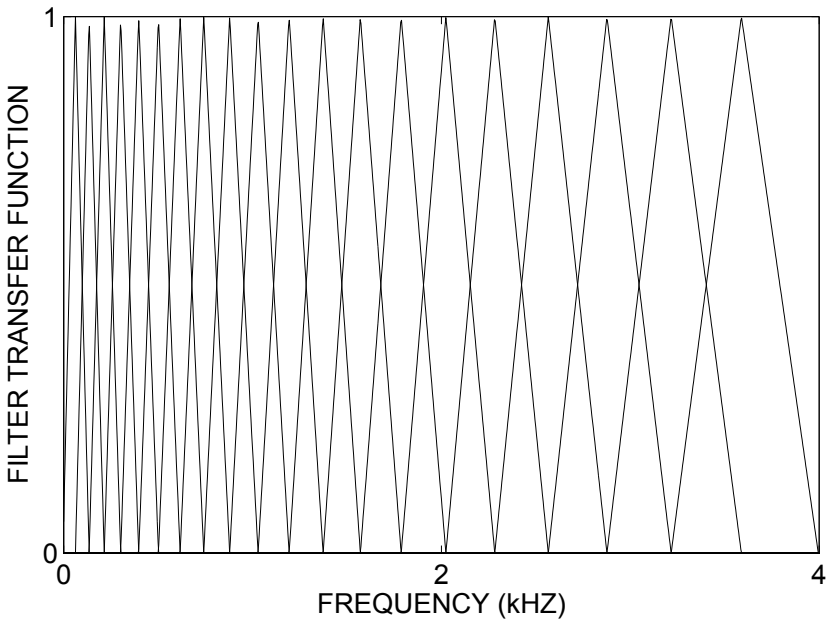

Figure 2.3. Triangular filters spaced evenly on the mel scale. The horizontal axis denotes ordinary, non-auditory frequency and the vertical axis denotes the magnitude of the filter transfer function.

Using these filters, logarithmic mel-filterbank energies for a signal frame whose DFT is S_k are given by

$$E_m = \log\left(\sum_{k=0}^{L} |S_k|^2 H_{m,k}\right). \quad (2.22)$$

MFCC analysis [27] is still the most widely adopted approach to speech and audio feature extraction [60], although some recent feature extraction solutions use the logarithmic mel-filterbank energies directly as input to neural networks. The stages of MFCC computation (Figure 2.4) are

1. estimation of the short-time magnitude spectrum,

2. computation of mel-filterbank energies using triangular bandpass filters in the frequency domain,

3. taking the logarithm of each filterbank output and

4. calculating the discrete cosine transformation (DCT) of the logarithmic filtered energies.

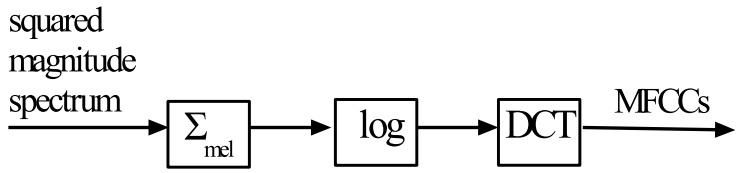

Figure 2.4. Data-flow diagram of the MFCC computation.

The DCT-II discrete cosine transform for M logarithmic mel-filterbank outputs E_m is given by [139] as

$$c_i = \sqrt{\frac{2}{M}} \sum_{j=1}^{M} E_m \cos\left(\frac{\pi i}{M}(j - 0.5)\right). \tag{2.23}$$

Comparing the steps of the MFCC processing chain with those of cepstrum analysis (Eq. 2.14), the MFCC is seen to be essentially a variation of the real cepstrum. It differs from the real cepstrum primarily in that logarithmic energies from an auditory filterbank are substituted for the ordinary logarithmic magnitude spectrum. Another difference between mel-cepstral and standard cepstral analysis is the use of the DCT-II in place of the IDFT to convert the logarithmic spectral representation to a cepstral representation.

Basic MFCC computation uses the DFT in the first step for magnitude spectrum estimation (essentially equivalent to the periodogram method of power spectrum estimation). In order to enhance the robustness of MFCC features in the presence of environmental noise, the spectrum estimation can be replaced

with another method. Simply replacing the DFT by LP, as done in some of the publications of this thesis, is sometimes found to be helpful; e.g., [28].

Besides MFCC analysis, other perceptual feature representations, such as *perceptual linear prediction* (PLP) [56], have been proposed especially in order to improve system robustness. These will be discussed in Section 4.4.2.

2.5 Long-term processing

In general, features computed using short-term analysis can only characterize the perceptual aspects of timbre, loudness and pitch (see Section 2.1.2). In order to represent the modulation frequency content of the speech signal, processing over longer time periods is necessary. As the (typically) frame-based short-time features are, nevertheless, suitable for modeling other important aspects, timbre in particular, long-term processing is often applied in cascade after initial analysis based on short-time frames.

A common approach to include information beyond the short-term frame is to concatenate short-term feature vectors with so-called *delta* (Δ) and *double-delta* ($\Delta\Delta$) features [44] (henceforth simply "deltas"). One way to compute them is to regress the feature x_n, where n is the frame index, linearly on an integer variable as [44]

$$\begin{aligned} \Delta_n &= \frac{\sum_{\theta=-W}^{W} \theta x_{n+\theta}}{\sum_{\theta=-W}^{W} \theta^2}, \\ \Delta\Delta_n &= \frac{\sum_{\theta=-W}^{W} \theta \Delta_{n+\theta}}{\sum_{\theta=-W}^{W} \theta^2}. \end{aligned} \tag{2.24}$$

The parameter W determines the width of the window used in computing the regression coefficients. In the studies of this thesis, a simpler approach is generally used, in which the deltas are obtained by simple differentiation:

$$\begin{aligned} \Delta_n &= x_{n+W} - x_{n-W} \\ \Delta\Delta_n &= \Delta_{n+W} - \Delta_{n+W}. \end{aligned} \tag{2.25}$$

Typically, the value of W is chosen as 1, meaning that the total time span covered by the Δ and $\Delta\Delta$ features are three and five frames, respectively.

Delta coefficients are a part of the typical frame-based feature vector, which consists of 39 features: the logarithmic energy (or an equivalent loudness-related parameter) and 12 MFCCs concatenated with the Δ and $\Delta\Delta$ coefficients of those 13 base features.

Despite taking into account information over neighboring frames also, deltas cannot capture information related to the lower modulation frequencies which are prominent in speech, e.g., down to 1 Hz (Section 2.1.2). Some long-term modeling approaches, such as feature filtering and cepstral normalization, are considered robustness-improving techniques and are discussed in Section 4. Sometimes, "complete" long-term representations of the modulation frequency characteristics of subbands are computed instead of relying on post-processing of higher-level short-term features. These representations include the *modulation spectra*, which in their simplest form can be implemented by frequency transforms applied to short-term spectral parameters [137], and *frequency-domain linear prediction* (FDLP) [6, 45], which generates minimum-phase temporal envelopes of signal subbands by means of linear predictive analysis applied to subband DCT sequences. These multivariate representations can be used for arbitrarily accurate representation of the long-term dynamics of each subband, but as the relevance of any individual variable may be limited, they are typically used as bases for lower-dimensional feature representations. In the general case, the most common approach for including modulation frequency information in applications such as paralinguistic analysis is to compute selected functionals of carefully chosen short-term features, or low-level descriptors (LLDs), typically on the utterance level [120]. Besides modeling low modulation frequencies, these features have the advantage that generic machine learning methods can be easily applied to their analysis. However, it is often not immediately clear which functionals to choose to represent speech utterances. Therefore, a long-term feature set often ends up being large and consisting of a diverse set of functionals applied to a comprehensive set of LLDs to characterize the short-term acoustics. With a large long-term feature set (often consisting of thousands of features), machine learning methods capable of tackling the effects of high dimensionality must be employed in order to obtain ideal results in a given application [121, 122, 123]. *Feature selection* methods, such as sequential forward selection [132], sequential backward elimination [83] and their "floating" modifications [106], are also often used when dealing with large feature sets. Publication VIII discusses feature selection methods.

3. Machine learning applications in speech processing

Machine learning is used in many tasks across the field of speech processing. The methods typically applied are briefly discussed in Section 3.1. This work focuses on applications where machine learning is in a primary role, discussed in Sections 3.2-3.5. In addition, application areas that use machine learning secondarily include, e.g., speech enhancement and speech synthesis.

3.1 Machine learning methods

Machine learning methods that are frequently used in speech analysis, and also used in the studies of this thesis, are briefly reviewed. The focus in this thesis is on classification, but examples of applying the methods to regression tasks are also mentioned.

Let $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ be an observed dataset of feature vectors, each with d elements. Let ω be a pattern of random variables, called the *state of nature* [34], that produces X. The random variables constituting ω are chosen based on the machine learning application. Examples of ω include speaker identity (in speaker recognition), a sequence of words in a given language (in speech recognition) or the age of the speaker (in paralinguistic analysis). The goal of the machine learning application is to infer the true value of the state of nature ω from the observed features X.

The natural probabilistic approach would be to find the inferred value $\hat{\omega}$ such that

$$\hat{\omega} = \underset{\omega}{\operatorname{argmax}}\, P(\omega|X). \tag{3.1}$$

However, *explicitly* modeling the probability distribution $P(\omega|X)$ in the case of continuous-valued features X is difficult. Even if these conditional distributions of ω were to be estimated in a limited number of vector quantization regions that together cover the space spanned by the vectors in X, the curse

of dimensionality would make this estimation from any practically feasible amount of training data inaccurate already with a relatively low dimensionality of X [34]. A common solution to this problem is the application of *Bayes' formula*:

$$P(\omega|X) = \frac{P(X|\omega)P(\omega)}{P(X)} \tag{3.2}$$

so that Eq. 3.1 becomes

$$\hat{\omega} = \underset{\omega}{\operatorname{argmax}} \frac{P(X|\omega)P(\omega)}{P(X)}. \tag{3.3}$$

Because $P(X)$ does not depend on ω, Eq. 3.3 obviously simplifies to

$$\hat{\omega} = \underset{\omega}{\operatorname{argmax}}\, P(X|\omega)P(\omega), \tag{3.4}$$

where $P(\omega)$ is the *prior distribution* of ω. It can be chosen as uniform, in which case it does not effect Eq. 3.4, or it can be chosen based on prior considerations about ω. In automatic speech recognition (ASR), $P(\omega)$ is determined by *language modeling*, an important component of an ASR system. $P(X|\omega)$ is the conditional density of the observed feature vector (or vector sequence) given the state of nature ω. Thus far, we have discussed ω in general terms, including the possibilities that it is a *continuous-valued* or an *ordered* variable. In these cases, the machine learning task would amount to *regression*. However, this thesis concentrates on automatic recognition of speech *classes* so that ω is effectively a *categorical* variable. In these applications, X typically consists of continuous-valued features. With these considerations, $P(X|\omega)$ is the *class-specific* probability density function (PDF) of an observed feature vector or feature vector sequence X.

A unimodal distribution, such as a (univariate or multivariate) Gaussian distribution, is frequently an inadequate model for representing probability distributions with complex shapes. However, an arbitrarily complex PDF shape can be modeled by using a *mixture* of a sufficiently large number of such distributions [113]. A *Gaussian mixture model* (GMM) with J *mixture components* is parametrized by $\lambda = \{P_1, \ldots, P_J, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_J, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_J\}$, where P_j are the mixture weights (component priors) and the pairs $(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ of $(d \times 1)$ mean vectors $\boldsymbol{\mu}_j$ and $(d \times d)$ covariance matrices $\boldsymbol{\Sigma}_j$ specify J multivariate Gaussian distributions. The PDF according to a GMM is given by

$$p(\boldsymbol{x}|\lambda) = \sum_{j=1}^{J} P_j b_j(\boldsymbol{x}) \tag{3.5}$$

with Gaussian component distributions

$$b_j(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_j|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)'\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right). \quad (3.6)$$

Prior constraints applied to the covariance matrices $\boldsymbol{\Sigma}_j$ have a large effect on the GMM. Each covariance matrix can be full or diagonal. In practice, diagonal covariance is typically applied because such GMMs are easier to train and often perform better than full-covariance GMMs [112]. In addition, if the amount of training data is very limited, diagonal-covariance GMMs can be trained with less data because they have less free parameters to estimate.

Each GMM component typically has its own dedicated covariance matrix, although the covariance matrix can also be shared among components. In the case of one global covariance matrix shared by each component, a full covariance structure may again become feasible, but its modeling capacity may not reach that of component-specific diagonal covariances matrices. In the following, the training data is denoted as $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$.

The parameters λ of a GMM can be iteratively trained using an application of the expectation-maximization (EM) principle [30]. For component-specific, diagonal covariance matrices, the re-estimation algorithm is [9]:

1. Either initialize λ and start from step 2 or initialize $\gamma_{n,i}$ and start from step 3.

2. Expectation (E) step: Compute the difference $\boldsymbol{e}_{n,i} = (e_{n,i}(1), \ldots, e_{n,i}(d))'$ of each vector to each of the component means:

$$\boldsymbol{e}_{n,i} = \boldsymbol{x}_n - \boldsymbol{\mu}_i, \qquad 1 \le n \le N, \quad 1 \le i \le J.$$

   Use $\boldsymbol{e}_{n,i}$ to compute posterior probability of the nth data vector being generated by the ith component:

$$\gamma_{n,i} = \frac{P_i b_i(\boldsymbol{x}_n)}{\sum_{j=1}^{J} P_j b_j(\boldsymbol{x}_n)} = \frac{P_i(1/\sqrt{(2\pi)^d|\boldsymbol{\Sigma}_i|})\exp\left(-\frac{1}{2}\boldsymbol{e}'_{n,i}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{e}_{n,i}\right)}{\sum_j^J P_j(1/\sqrt{(2\pi)^d|\boldsymbol{\Sigma}_j|})\exp\left(-\frac{1}{2}\boldsymbol{e}'_{n,j}\boldsymbol{\Sigma}_j^{-1}\boldsymbol{e}_{n,j}\right)}.$$

3. Maximization (M) step: Re-estimate mixture weights:

$$p_i = \frac{1}{N}\sum_{n=1}^{N}\gamma_{n,i}, \qquad 1 \le i \le J.$$

   Re-estimate component mean vectors:

$$\boldsymbol{\mu}_i = \frac{\sum_{n=1}^{N}\gamma_{n,i}\boldsymbol{x}_n}{\sum_{n=1}^{N}\gamma_{n,i}}, \qquad 1 \le i \le J.$$

Re-estimate the variance parameters for the diagonal covariance matrices:

$$\sigma_{i,j}^2 = \frac{\sum_{n=1}^{N} \gamma_{n,i} e_{n,i}^2(j)}{\sum_{n=1}^{N} \gamma_{n,i}}, \qquad 1 \leq i \leq J, \quad 1 \leq j \leq d.$$

(Note that in some variants of GMM re-estimation, e.g., [113], $\boldsymbol{e}_{n,i}$ are re-computed using the newly estimated mean vectors $\boldsymbol{\mu}_i$ before estimating the variances $\sigma_{i,j}^2$).

4. If a convergence criterion is met or a specified number of iterations has been run, exit. Otherwise, go to step 2.

Any EM algorithm is guaranteed to converge at least towards a local, if not the global, likelihood maximum, because each iteration is guaranteed to increase the model likelihood $L(\lambda|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ [30].

In using GMMs for supervised classification of the sequence $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$, with $\omega \in \{\omega_0, \ldots, \omega_{C-1}\}$, logarithmic posterior probabilities are typically obtained and averaged over the $M = |X|$ vectors as $(1/M)\log(P(X|\omega = \omega_k)) = (1/M)\sum_{n=1}^{M} \log(p(\boldsymbol{x}_n|\lambda_k)) = (1/M)\sum_{n=1}^{M} \log(\sum_{j=1}^{J} P_{j,k} b_{j,k}(\boldsymbol{x}_n))$. Eq. 3.4 can be used to select the most likely class. In *detection* applications where $C = 2$, a *logarithmic likelihood ratio test* is typically used for averaged logarithmic likelihoods [112]: class ω_1 is decided if $\log(P(X|\omega = \omega_1))/M - \log(P(X|\omega = \omega_0))/M > T$, where T is an adjustable threshold value. The choice of T gives rise to different values of the *miss rate* p_{miss} (the rate of failing to detect class ω_1) and the *false alarm rate* p_{fa} (the rate of mistakenly detecting class ω_1) over an evaluation data set. A *detection-error-tradeoff* (DET) curve can be plotted to show p_{miss} against p_{fa} over these different operating points [85]. The older *receiver operating characteristics* (ROC) curve shows the same information as a DET curve but plots $1 - p_{\text{miss}}$ against p_{fa} without using logarithmic axes like the DET plot. Both curves can be plotted by increasing the detection threshold T such that p_{fa} decreases from 1 to 0 and p_{miss} increases from 0 to 1. The *equal error rate* (EER) is obtained as $p_{\text{miss}} = p_{\text{fa}}$ at the corresponding value of T.

The popular *k-means* algorithm, which is used both for clustering and for generating vector quantization (VQ) codebooks [63], is identified by Bottou and Bengio as an *EM-style* iterative algorithm [13]. In a sense, k-means is a deterministic version of Gaussian mixture learning. The VQ model learned by k-means is parametrized only by the cluster mean vectors as $\lambda = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_J\}$. The similarity between the two algorithms can be noticed by considering the following formulation of k-means:

1. Either initialize λ and start from step 2 or initialize $g_{n,i}$ and start from step 3.

2. Quantization step: Compute the difference $\boldsymbol{e}_{n,i} = (e_{n,i}(1), \ldots, e_{n,i}(d))'$ of each vector to each of the component means:

$$\boldsymbol{e}_{n,i} = \boldsymbol{x}_n - \boldsymbol{\mu}_i, \qquad 1 \leq n \leq N, \quad 1 \leq i \leq J.$$

   Use the $\boldsymbol{e}_{n,i}$ to assign (quantize) each of the data vectors into the nearest of the J clusters (squared Euclidean distance):

$$g_{n,i} = \begin{cases} 1 & : i = \text{argmin}_j(\boldsymbol{e}'_{n,j}\boldsymbol{e}_{n,j}) \\ 0 & : \text{otherwise} \end{cases}$$

3. Re-estimation step:

   Re-estimate component mean vectors:

$$\boldsymbol{\mu}_i = \frac{\sum_{n=1}^{N} g_{n,i}\boldsymbol{x}_n}{\sum_{n=1}^{N} g_{n,i}}, \qquad 1 \leq i \leq J.$$

4. If a convergence criterion is met or a specified number of iterations has been run, exit. Otherwise, go to step 2.

Each iteration of k-means is guaranteed to decrease the total quantization error $\sum_{n=1}^{N}\sum_{i=1}^{J} g_{n,i}\boldsymbol{e}'_{n,i}\boldsymbol{e}_{n,i}$ and, therefore, to converge towards its local minimum.

In both GMM estimation and k-means, the fact that the likelihood can only increase constrains the directions that parameters can change towards during the iteration. Therefore, both methods converge towards a local optimum determined by the initial parameter set and are sensitive to initial parameter values. In practice, it has been noticed that GMM re-estimation is not particularly sensitive to the initial values of the weights P_i or the covariance matrices $\boldsymbol{\Sigma}_i$ as long as they are initialized reasonably (positive weights and sufficiently small, positive variances). Initialization of the mean vectors is often noticed to be more critical, and indeed, many initialization methods have been proposed for k-means [2, 14, 55, 68, 98]. The mean vectors in GMM re-estimation can be initialized by using the same methods, or by using the result of the k-means algorithm after its convergence. Katsavounidis, Kuo and Zhang [68] proposed a simple, deterministic method for the initialization of k-means (also known as Lloyd iteration):

1. Choose the vector with the maximum Euclidean norm $\|\boldsymbol{x}_n\| = \boldsymbol{x}'_n\boldsymbol{x}_n$ as the

first cluster center:

$$\boldsymbol{\mu}_1 = \underset{\boldsymbol{x}_n}{\operatorname{argmax}} \|\boldsymbol{x}_n\|.$$

2. For $i = 2$ to J:

   Choose the vector with the greatest distance from the current codebook as the new center:

$$\boldsymbol{\mu}_i = \underset{\boldsymbol{x}_n}{\operatorname{argmax}} \|\boldsymbol{x}_n - \underset{\boldsymbol{\mu}_j, 1 \leq j \leq i-1}{\operatorname{argmin}} \left(\|\boldsymbol{x}_n - \boldsymbol{\mu}_j\|\right)\|.$$

This method is sometimes termed the KKZ algorithm, according to the initials of the authors (e.g., [55]). The first cluster center selected is the one with the maximum norm within the data set and the second center selected is the one furthest away from it. Therefore, this method tends to select initial points near the outer boundaries of the d-dimensional point cloud represented by the data set $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$. During the subsequent EM or k-means iteration, these center points typically converge towards more central locations with respect to the point cloud, as illustrated in Figure 3.1.

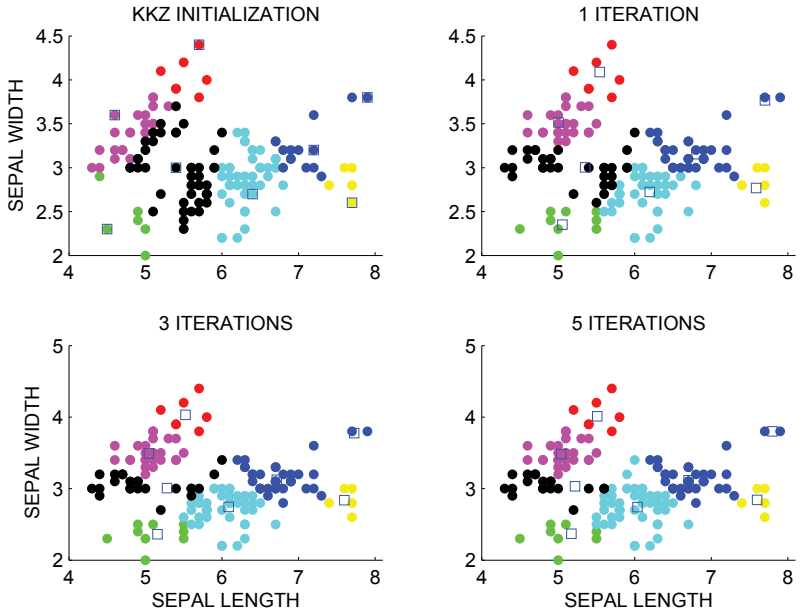

Figure 3.1. Illustration of the EM algorithm to train an eight-component Gaussian mixture model, with mean vectors initialized by the KKZ algorithm, applied to the sepal length and sepal width measurements in Fisher's iris data set [42]. The mean vectors at each iteration stage are shown as squares. The colors indicate the component-specific cluster to which each point belongs with the highest posterior probability $\gamma_{n,i}$.

Bayesian classification according to Eq. 3.4 requires explicit modeling of class-specific observation PDFs $P(X|\omega = \omega_j)$. *Nearest-neighbor classification* is an alternative approach that avoids the difficulty of estimating explicit PDF models, or any kind of models. Starting from the basic problem in Eq. 3.1, *k-nearest-neighbors* (kNN) classification [10, 34, 127] estimates the conditional class distribution $P(\omega|X)$ by (1) finding in the training data the k feature vectors or sequences of feature vectors that are closest to the test input X (based on some metric), (2) calculating the frequencies of the different classes among these k nearest neighbors and (3) choosing the mode, or most frequent value, of that distribution as the predicted class label.

kNN classification counters the most obvious effect of the curse of dimensionality – that the high-dimensional feature space is mostly empty of training samples–by centering distribution modeling specifically around each test input X and requiring the neighborhood to contain at least a predefined number (k) of training points. This enables using Eq. 3.1 directly to make class decisions. When the discriminative nature of the features is high, kNN is a powerful, non-linear classification method. However, each feature contributes to the size of the k-neighborhood and irrelevant features may bring irrelevant points within the k-neighborhood. This biases the estimation of $P(\omega|X)$ based on k samples. kNN classification is thus subject to the curse of dimensionality in this respect. In Publication VIII, kNN is used as the classifier in evaluating feature selection methods.

While kNN is most often used for classification, it is also applicable to regression, e.g., [111]. In this case, the regression result is obtained simply as the average of the dependent variable among those k observations for which the predictor variables are closest to the test input X.

The support vector machine (SVM) is a popular machine learning method [4, 10, 18]. It can be used for both classification and regression problems. For the purpose of binary classification, when the data is linearly separable, the aim is to find the separating hyperplane that maximizes the distance (the *margin*) between the hyperplane and the correctly classified class instances closest to it on either side (the *support vectors*). In classification problems that are not linearly separable, *slack variables* and a misclassification cost tradeoff factor can be added in the margin optimization to allow misclassifications (the *soft margin* method). Alternatively, using a suitable basis or *kernel* function, the feature vectors can be implicitly mapped into a higher-dimensional space in which linear separation of the classes in the training set becomes possible.

Classification and regression trees (CART) [17] is a method for learning de-

cision trees for either classification or regression. It builds a binary decision tree which always splits the data by thresholding one of the features, leading to a rectangular partition of the feature space. In training, the choice of the feature and threshold at each node is determined by the *purity* of the two subsets generated by the split operation. *Random forests* [16] is a related method which uses multiple decision trees, each typically based on a randomly selected subset of the features [61]. Similarly to SVMs, random forests are often found to be robust against irrelevant or noisy features in high-dimensional problems. They both appear as baseline methods in Publication VIII.

Recently, *artificial neural networks* [87] have re-emerged as a prominent pattern analysis and modeling method in the form of *deep neural networks* (DNNs). These networks are motivated by computational modeling of the central nervous system. They consist of layers of neurons, each of which computes an output activation using a nonlinear, monotone activation function of a linear combination of its inputs. Successive layers of conventional *feed-forward* networks are fully connected in the sense that each of the outputs of the preceding layer act as inputs to each of the neurons of the following layer. This way, the network generates a new kind of multidimensional abstraction of its multidimensional input in every layer. Modern DNNs are called "deep" because of having at least two hidden layers (between the input and output layers) and generally more hidden layers than the earlier networks which were used in the 1980s and 1990s. Advancements in computation and data storage technologies, as well as in training methods, have permitted training and deployment of these larger networks, which can generate more complex abstractions and thus solve more complex practical problems, as evidenced by their recent success in ASR applications [60]. DNNs can also be *recurrent*, i.e., contain feedback connections, so that they can model long-term temporal dependencies. DNNs can be considered a "black box" as well as a "brute force" learning method which, given sufficient training data, can learn to model various complex dependencies between the input and output variables.

The Euclidean distance is a typical metric to use in nearest-neighbor methods. When X is a sequence of vectors, *dynamic time warping* (DTW) can be applied instead [92]. DTW is a dynamic programming technique that finds a minimum-cost path through a $N_1 \times N_2$ grid, where N_1 and N_2 are the lengths of the two sequences. The cost associated with one point of the grid is a measure of dissimilarity between the corresponding elements of the two sequences. The result of DTW is an alignment between the two sequences that is optimal according to the chosen constraints. Minimized costs can be used as

distance measures between sequences in nearest-neighbor methods, as is done in Publication I.

Hidden Markov models (HMMs) are stochastic models for the PDFs of feature vector sequences [65, 108]. They can be viewed as containing elements from both Gaussian mixture modeling and dynamic time warping. Similarly to GMMs, HMMs are specified by discrete probability parameters that characterize the prior distributions of a hidden variable as well as the (typically continuous) observation distribution parameters associated with each state or component. HMMs are typically trained using an implementation of the EM principle known as the *Baum-Welch algorithm*. Similarly to DTW, an optimal path through a *trellis* of the states of a HMM can be found using an implementation of dynamic programming (the *Viterbi algorithm*). HMMs are extensively used in large-vocabulary continuous speech recognition. Like DNNs, they are not used in the studies of this thesis but are mentioned for completeness due to their importance in both ASR applications and statistical parametric speech synthesis.

3.2 Speech event detection

Speech event detection can be viewed as a sub-category of audio event detection [52], where the target class of detection is some type of vocal activity. Examples include detection of speech in general; shouts [115] (as discussed in Publication III); screams [129, 130]; and non-neutral speech [110]. On a time scale in the order of seconds, these detection tasks have applications in audio-based automatic surveillance and acoustic environment monitoring [52, 110, 115, 129, 130]. Similarly to paralinguistic analysis (Section 3.4), long-term speech event detection may also be used to assist speech and speaker recognition systems in choosing suitable acoustic models or in assisting speech-based user interfaces to be aware of the speaking situation. The latter applications can be related to a system with microphones listening to its environment or a telephone-based system (typically in a call center). On a related note, when detection of speech activity is done on a fine time scale, such as the short-time frame level, it is commonly referred to as voice activity detection (VAD) [125] and is an important component that saves computational resources by assisting speech coding, speech recognition and speaker recognition systems to only process meaningful speech frames.

A typical approach in speech event detection is to start with short-time timbral features, such as MFCCs [110, 115], and to apply a generic pattern

recognition method, such as GMM-based Bayesian classification [110, 115, 129, 130] or SVM [115]. However, detection of a more specific target speech class with its typical characteristics, such as angry speech [19, 40, 99], has typically been found to benefit from long-term modeling of the modulation frequency content in order to separate the target class from other types of speech. On the other hand, such applications approach paralinguistic analysis (Section 3.4).

3.3 Speaker recognition

Supervised speaker recognition can be divided into *speaker identification*, where the goal is to identify the speaker among a closed speaker population, and *speaker verification* (or speaker detection [112]), where the goal is to detect whether the speaker is a known target speaker or someone else. Both approaches can be text-dependent or text-independent, depending on whether the user is required to speak a given phrase or not. Publications II and VI focus on text-independent speaker verification.

Speaker verification has been a topic of active research and certain paradigms have emerged over the years. The Gaussian mixture model-universal background model (GMM-UBM), by Reynolds et al. [112], has been widely used. It consists of first training a *universal background model* (UBM), a large GMM with typically hundreds or some thousands of components, to parametrize the complete background speaker population. Speaker-specific GMMs are obtained by adapting the UBM with speaker-specific training data. In the detection phase, a logarithmic likelihood ratio test, for logarithmic likelihoods averaged over the frames of an utterance, is performed between the speaker-specific model and the UBM in order to accept or reject the test speaker as the target speaker. SVM-based classification of *GMM supervectors*, formed by concatenating the mean vectors of each GMM component, is a more recent approach which gives good performance with nuisance attribute projection (NAP) [21]. Recently, Dehak et al. proposed a factor-analysis-based front-end [29]. The output representation, commonly referred to as *i-vectors*, generally outperforms the previous approaches and is currently considered to be the state-of-the-art in speaker verification.

3.4 Paralinguistic speech analysis

Paralinguistics refers to the study of information carried by speech *alongside* (*para-*) its linguistic content [120]. *Computational paralinguistics* refers to computerized analysis of paralinguistic aspects of the speech signal. Obviously, speech recognition does not belong to this field, as it is exclusively concerned *with* the linguistic content. Typically, neither the detection of speech events nor speaker recognition are considered to be essentially paralinguistic tasks, as they can be viewed to be concerned with the *speaking situation* rather than the information conveyed by the speech. These constraints leave the following types of information to be uncovered by computational paralinguistic analysis of speech (see also [119, 120]):

- Short-term/instantaneous: e.g., speaking mode/vocal effort (normal, shouted, whispered) [90, 102, 142]; voice quality (normal, creaky, breathy) [33]; detection of non-speech events such as laughter [123]; melody transcription in singing or humming [46, 104].

- Medium-term *states*: e.g., emotional state (e.g., happy, sad, angry, surprised) [36, 75, 89, 118, 136]; affective dimensions (activation, valence and dominance) [20, 131, 135]; depression [80]; physical stress [47, 51, 97]; sleepiness [74]; intoxication [12].

- Long-term characteristics and *traits*: e.g., age and gender [77]; height [133]; various pathologies [100, 111]; likability of voice [100]; personality traits [100].

Collectively, the long-term traits have an evident connection to speaker recognition [120]: in theory, reliably identifying a sufficient number of such variables permits us to single out a given speaker from an arbitrarily large speaker population. More generally, the inferred variables generated by paralinguistic analyses can themselves serve as high-level features for other problems. For example, analysis of vocal effort could serve as one feature extractor in an emotion detection system, or paralinguistic features could influence acoustic modeling in the decoding phase of automatic speech recognition (Section 3.5).

Many of the above listed paralinguistic characteristics of speech can variably be expressed as either continuous-valued or ordered categorical variables. In

many occasions, such as in the Computational Paralinguistics Challenges of the Interspeech conference [121, 122, 123], binary categorical variables are considered as targets of prediction, often on a scale of "low" and "high" or "present" and "not present".

As discussed in Section 2.1.2, the basic components of auditory perception are loudness, pitch, timbre and subjective duration. Most speech feature extraction methods are based on timbre and have been originally developed for automatic speech recognition, but are also applicable to speaker recognition. This is because timbre is a good discriminator of speech sounds (in non-tonal languages) as well as of speaker-specific vocal tract configurations. In paralinguistics, timbral features are frequently not sufficient, however. Depending on the exact task, pitch and subjective duration of component sounds can also play an important role (loudness, however, can be reduced to a combination of absolute sound level, which is generally not useful, and timbral aspects [143]).

Pitch perception is connected to the fundamental frequency F0, which can be explicitly modeled by *pitch estimation* algorithms [59] and also manifests itself in the harmonic fine structure of the spectrum. Subjective duration leads to the perception of rhythm. It is closely associated with the modulation frequency content of the signal (in the physical domain) or its auditory representations in the perceptual domain.

Given that different paralinguistic machine learning problems may require quite different feature representations, it is probably natural that more than a few different types of systems are being employed. Most systems can be roughly organized into two major categories.

The first main approach is based on timbral features and statistical pattern recognition but includes pitch and modulation aspects as needed, in some form. For example, the timbral feature vector can be appended with additional features related to pitch or modulation frequencies [89], or class-specific modulation frequency characteristics can be modeled by explicit modulation filtering [99].

The second, feature-intensive approach does not assume to know exactly the types of features that would be useful for the given task. Instead, a large number of long-term functionals of short-term acoustic parameters are computed (in a sense, a brute-force approach facilitated by increased computational resources), leading to a high-dimensional feature space. This initial feature set can be designed to be very comprehensive [121, 122, 123] or to focus on some timbral and some modulation-related features [25, 136]. As discussed earlier, in machine learning, the curse of dimensionality [34] means that high-dimensional

feature spaces are sparsely populated by limited training data since the number of data points inside a volume unit decreases with an increasing feature space dimensionality. One effect of this is to weaken the statistical reliability of trained systems. Therefore, the machine learning methods applied by the feature-intensive approach must be resistant to this aspect of the curse of dimensionality. Methods that meet this requirement include SVM and random forests [122] as well as feature selection and dimensionality reduction methods capable of working with high-dimensional feature spaces [100, 111].

In this thesis, Publications IV, VII and VIII belong to the field of paralinguistics while publication III has a connection to the detection of the speaking mode, even though its immediate application is in the field of speech event detection.

3.5 Automatic speech recognition

Automatic speech recognition (ASR) is a field of speech technology that has been extensively studied since the 1960s. ASR has found its first practical applications with constrained speech inputs, such as in control tasks (simple speech commands) and dictation (speech-to-text transcription). In recent years, the prominence of ASR has further increased with, e.g., the advent of workable speech-based user interfaces and conversational search agents in smartphones. In addition to relaxing the domain constraints of these applications, projected future applications of ASR (not yet in wide use at the time of this writing) include speech-to-speech translation and reliable speech indexing for searching large multimedia archives with greatly varying content. As the real-world applications target less and less constrained speech material, robustness with respect to natural speech signal variability becomes increasingly important.

In comparison to the other speech-related machine learning applications discussed above, the ASR problem is clearly the most complex. In applications such as text-independent speaker recognition, speech detection and speaking mode analysis, it often suffices to model the statistics of short-time timbral features. Additional modeling of long-term rhythmic patterns of loudness, timbre and pitch may be required in the more complex paralinguistic analyses. ASR, however, requires the modeling of sequences of word tokens consisting of timbral variations on a fine time scale; in the long term, this task is still usually accomplished by finding (*decoding*) an optimal path through a network of HMMs [65], while modeling of short-term acoustic properties with

GMMs has recently been mostly replaced by DNNs. Still, despite the already higher complexity in comparison to the other applications, the performance of ASR cannot rely simply on accurate *acoustic modeling*, regardless of the time scale. Information about the language being spoken is also required. For this purpose, ASR systems employ *language models* that model the probability distribution of linguistic sequences. The importance of language models can be understood by considering speech recognition by humans: it is very difficult to accurately transcribe text spoken in a strange language (even with knowledge of the phoneme space of that language). Current language models are often low-order n-grams, which can provide good performance in limited linguistic domains (e.g., broadcast news transcription and voice search applications). However, it can be argued that understanding the spoken message is also fundamental to robust and accurate speech recognition that will match human performance. For example, humans can often guess the content of the next sentence based on long-term context (perhaps also applying paralinguistic and visual cues when applicable), which is arguably more than what can be accomplished by simple language models that consider short word sequences.

The contribution of this thesis to ASR is twofold. Firstly, the methods proposed for short-time feature extraction and long-term feature post-processing can find application in improving the robustness of ASR systems. In addition to the word recognition experiments in Publication I, additive-noise robustness of large-vocabulary continuous speech recognition (LVCSR) systems has been improved by time-weighted linear predictive methods in recent studies [69, 101]. In [69], LVCSR performance is improved using a special case of the generic XLP formulation described in Publication IV. This special case was originally proposed for robust speaker verification in [103]. Another contribution of this thesis to ASR is to provide paralinguistic analysis systems for detecting different speaking styles, such as various emotional states (Publications IV and VII) or speaking with high vocal effort (Publication III). Such analyses can allow ASR systems to choose or adapt the acoustic models in the decoding phase according to the current speaking style [8, 140, 141].

4. Robustness

This section discusses signal variability, its contribution to mismatch in machine learning systems and robustness-improving techniques designed to counteract the mismatch.

4.1 Causes of signal variability

4.1.1 Additive noise

The desired clean signal s_n is typically corrupted by additive noise v_n, i.e.,

$$y_n = s_n + v_n. \tag{4.1}$$

Ambient sounds of the recording environment as well as effects of the recording or transmission channel can cause additive noise to be present in the speech signal. The noise may make it more difficult to discern the acoustic properties of the original signal. The way in which additive noise affects recognition task performance is determined by the interaction of the spectral/temporal characteristics of the additive noise and the class-separating cues [79].

4.1.2 Effect of the channel

Broadly, the *channel* comprises the acoustical transfer function of the recording environment (room response and echoes) and the acoustical and electrical transfer functions of the recording equipment and the actual transmission channel. Each of these effects causes a convolutive distortion which can be modeled in the time domain as

$$y_n = s_n * g_n \tag{4.2}$$

and, equivalently [91], in the frequency domain as

$$Y(z) = S(z)G(z). \tag{4.3}$$

4.1.3 Source variability

In the context of speech processing, source variability–signal variability due to different representations of semantically similar information–refers to acoustic variation caused by speaker-related effects. These include fundamental frequency patterns, vocal effort, speaking rate, speaking styles and transient emotional or physical states of the speaker, among other things. They affect, in different ways, the different aspects of auditory perception: timbre, loudness, pitch and subjective duration (Section 2.1.2) via their physical counterparts of the short-time spectrum, intensity, fundamental frequency and rhythm or duration. It can be noticed that these speaker-related causes of variability are also among the typical recognition targets in paralinguistic analysis (Section 3.4).

4.2 Mismatch

In machine learning applications of audio signal processing, the goal is to infer "hidden" information based on the observed signal. The signal characteristics are automatically learned in the training phase from features extracted from the training signals. The training material depicts only a certain subset of all possible conditions. In this thesis, *condition* is used informally to refer to a certain configuration of additive noise, recording channel characteristics, speaking style, speaker state and speaker characteristics. If the conditions between training and actual usage (test phase) of the system are different so as to have potential influence on the statistics of typical features, *mismatch* is said to be present. Mismatch may result in differences between the feature statistics learned in the training phase and those encountered in the test phase, which in turn may result in performance degradation of the machine learning system. To the degree that the system is able to maintain its performance level despite condition mismatch, it is said to be *robust*.

If the signal to be analyzed is affected by any kind of degradation or distortion, such as additive background noise or a degraded convolutive channel, the expected amount of mismatch increases because it becomes less likely that this particular type of distortion has been covered by the training material. In speech signal processing, mismatch can also arise simply because

similar information content can be expressed in many different forms due to speaker-related effects such as voice quality, speaking style, fundamental frequency and vocal effort, as discussed in Section 4.1.3. Vocal effort mismatch, for example, is found to decrease the performance in speech [140, 141] and speaker recognition [124, 142]. *Multi-condition training*, which does try to account for all different conditions already in the training material, is often found to improve the results of machine learning systems. In the test phase, a multi-condition-trained machine learning system may benefit from paralinguistic analyses (Section 3.4) to help it choose the correct acoustical models. However, the problem of multi-condition training lies in the difficulty of obtaining sufficient training data to comprehensively cover each possible usage condition. Therefore, robustness-improving signal processing techniques often have to be used.

Considering that performance degradation is caused by differences in training and test feature statistics (which in turn are caused by condition mismatch), different approaches can be taken to improve the system's robustness, i.e., to prevent performance degradation due to mismatch. The choice of the approach depends primarily on (1) which parts of the process chain of a given machine learning system it is easiest to work on (Figure 1.1), (2) the type of variability that is causing the mismatch and (3) the available techniques. In the preprocessing step before feature extraction (Figure 1.1), *speech enhancement* operations can be used to reduce the amount of degradation present in the signal (Section 4.3). In the feature extraction phase, on the short-time level, *robust spectrum analysis* methods can be used to focus on information less affected by signal corruption (Section 4.4.1). Another popular approach on the short-time feature level is to improve the modeling of auditory perception in feature extraction, as the human auditory system is known to be fairly robust (Section 4.4.2). On a longer time scale, spanning many short-time frames, filtering methods can be applied to feature vector sequences (Section 4.4.3), or the primary features can themselves be long-term functions of the signal. These long-term processing methods may result in improved modeling of aspects of neural auditory processing, such as modulation-frequency selectivity (Section 2.1). Finally, many more methods to improve robustness are available on the level of the classification or regression sub-system (Figure 1.1). These include discriminative training [107], missing data imputation [26] and using an informative, typically high-dimensional feature set in combination with a suitable classification or regression method [60, 119].

The focus of this thesis is on improving robustness on the feature level by

emphasizing relevant information in different ways. The majority of the studies, Publications I to VI, are concerned with robust spectrum analysis based on temporal weighting of the information within a short-time analysis frame (Section 4.4.1). Publication VII deals with feature post-processing to emphasize relevant modulation frequencies (Section 4.4.3), and Publication VIII studies the automatic selection of long-term features in the high-dimensional classification approach.

4.3 Speech enhancement

Speech enhancement is a collective name for signal processing operations that are applied to the digital speech signal in order to improve one or more of the following aspects:

- Subjective quality (as judged by a human listener)
- Intelligibility (as judged by a human listener)
- Automatic analysis accuracy, e.g., in one of the problems discussed in Section 3

In the context of this thesis, speech enhancement is considered as a robustness-improving preprocessing step that can optionally be applied to the signal before the feature extraction phase in order to reduce possible degradations. In addition, as speech enhancement and robust feature extraction share partially similar goals and methods–both aim to reduce the effects of signal corruption in some representation domain–it is of interest to discuss certain prevalent approaches to speech enhancement.

Automatic removal of additive noise is a central speech enhancement task for which many methods have been developed [79]. Since the original *spectral subtraction* noise reduction algorithm, proposed by Boll in 1979 [11], many variants of speech enhancement algorithms have been proposed. Most of these algorithms utilize a training segment of the noise in order to construct an enhancement filter to remove the noise in separate short-term analysis frames and then use the overlap-add method [109] to resynthesize the modified signal. Two common classes of speech enhancement algorithms are the spectral subtraction algorithms, where the noise model is deterministic, and the *Wiener filter* algorithms, which can be realized using stochastic noise models.

Spectral subtraction [11] is a classical enhancement method for removing additive noise. Taking the z transform of both sides of Eq. 4.1 [91] gives

$$Y(z) = S(z) + V(z), \tag{4.4}$$

where $Y(z)$, $S(z)$ and $V(z)$ are the z transforms of the noise-corrupted signal, the clean signal and the additive noise, respectively. The basic idea of spectral subtraction is to estimate $S(z)$ (and thereby the clean signal s_n itself) so that the magnitude of the estimator is given by

$$|\hat{S}(z)|^c = |Y(z)|^c - N(z), \tag{4.5}$$

where $c > 0$ is an exponent and $N(z)$ is an averaged spectrum of the noise power $|V(z)|^c$,

$$N(z) = E\{|V(z)|^c\}, \tag{4.6}$$

which is to be estimated from parts of the signal known to contain only background noise. The phase of the spectral subtraction estimator $\hat{S}(z)$ is given directly by the phase of the noisy input signal $Y(z)$.

If the problem is cast as that of filtering the noise-corrupted signal y_n to obtain $\hat{s}_n$, i.e., as $\hat{S}(z) = H(z)Y(z)$, a filter $H(z)$ that realizes the above specifications is given by

$$H(z) = \frac{(|Y(z)|^c - N(z))^{1/c}}{|Y(z)|}, \tag{4.7}$$

which in [11] is used with $c = 1$, yielding $H(z) = 1 - \frac{N(z)}{|Y(z)|}$. The filtering operation then becomes equivalent to $\hat{S}(z) = (|Y(z)|^c - N(z))^{1/c} Y(z)/|Y(z)|$, meaning that the phase spectrum of $Y(z)$ is multiplied by the cth root of the result of spectral subtraction $|Y(z)|^c - N(z)$. There may be frequencies for which $N(z) > |Y(z)|^c$. Such noisy frequencies are usually assumed to be unrecoverable. Thus, it is enough to ensure that the subtraction result stays nonnegative. This can be accomplished by using half-wave rectified filtering $\hat{S}(z) = H_r(z)Y(z)$, where [11]

$$H_r(z) = \frac{H(z) + |H(z)|}{2}. \tag{4.8}$$

Random residual spikes that remain in non-active temporal and spectral regions after subtraction give rise to tonal noise (so-called musical noise). In [11], it is suggested to reduce musical noise by using a filtering operation where the magnitude of a given DFT frequency bin is replaced by its minimum value

chosen from the adjacent analysis frames, when the original magnitude falls below the maximum noise residual calculated during non-speech activity.

The above described basic principles of spectral subtraction follow the original work of Boll [11], who used $c = 1$ (*magnitude spectral subtraction*). Another common alternative is *power spectral subtraction* with $c = 2$ [63, 79]. The noise estimate, Eq. 4.6, is obtained during a non-speech segment and may be updated based on the decision of a voice activity detector in order to adapt to changes in the noise environment [79]. In Publication II, power spectral subtraction is used as preprocessing for a speaker verification system operating in noisy conditions.

Wiener filtering is another popular approach for noise removal. The mean squared error of the filtered target signal is minimized [54]. In the frequency domain, the *noncausal Wiener filter* is obtained based on the power spectra of both the desired signal and the noise, which must be known. This kind of filter is given by [63, 79]

$$H(z) = \frac{|S(z)|^2}{|S(z)|^2 + |V(z)|^2}. \tag{4.9}$$

In practice, this Wiener filter cannot be realized as the power spectra of both the desired signal s_n and the noise v_n would have to be known, and at least the former spectrum is unknown due to the very nature of the noise reduction problem. In order to work around this problem, iterative solutions which base new estimates of the Wiener filter on the enhanced signal obtained by the previous iteration's Wiener filter estimate can be used as first described by Lim and Oppenheim [78, 79]. Hansen and Clements [50] proposed a constrained approach to iterative Wiener filtering which imposes, in each iteration, across-time and across-iterations constraints affecting the newly estimated Wiener filter. Non-iterative approaches based on different types of *a priori* SNR estimation have also been developed for implementing Wiener filtering [62, 79, 116].

The minimum mean square error (MMSE) [38] and log-MMSE estimators [39], proposed by Ephraim and Malah in 1984 and 1985, respectively, are also efficient speech enhancement methods, whose performance still remains among the best of the published methods [94]. In the form that these methods were introduced, they involve the *decision-directed* estimation approach [38], which bases the spectral estimate of each frame partially on the estimates from previous frames via the *a priori* SNR estimate updated by using a memory coefficient. This can be viewed as one way of utilizing long-term information for speech enhancement. It has been argued that the absence of musical noise observed with the Ephraim and Malah noise suppression methods is primarily

due to the decision-directed approach [23, 116]. Scalart and Filho [116] applied the decision-directed estimation of *a priori* SNR also to implementing other methods, such as power spectral subtraction and Wiener filtering, and observed reduced musical noise. More recently, long-term modulation domains have been increasingly utilized in speech enhancement [73, 94, 96, 117]. Recently, spectral subtraction [96] and the decision-directed MMSE [94] methods have been applied in the spectral modulation domain, i.e., separately for each short-time frequency component across frames, and the latter method has been found to better handle speech nonstationarity by being able to utilize shorter long-term analysis windows to arrive in the spectral modulation domain.

4.4 Robust feature extraction

Some feature extraction approaches place particular emphasis on robustness. That is, the performance of a speech processing system using these methods is expected to degrade less in the presence of distortions and/or mismatch than when using a standard feature set.

On the short-time frame level, two main approaches to robust feature extraction can be distinguished. The "signal processing approach" relies on signal processing methods without additional perceptual considerations. For example, standard DFT spectrum analysis can be replaced with another spectrum analysis method that exhibits better robustness performance in the application domain. These non-perceptual methods may involve implicit assumptions that improve their performance with certain types of signals. The spectrum analysis methods studied in Publications I to VI fall into this category. Their common background is reviewed in Section 4.4.1.

An alternative approach to improving robustness is by means of improved perceptual modeling. The "perceptual approach" is justified by the notion that it is difficult to go very wrong in terms of robustness by modeling the human auditory perception, which is known to be quite robust [57]. However, high-level neural processing plays an important role in perception and, due to a lack of knowledge, it is more difficult to model than peripheral processing. This lack of psychoacoustical knowledge places practical boundaries on robustness gains obtainable by using perceptual models, which are largely confined to modeling the peripheral processing. While the popular MFCC feature extraction is already perceptually motivated–it employs spectrum analysis on a nonuniform frequency scale, similarly to the basilar membrane, as well as a compressed (logarithmic) dynamic range–it is not a complete model of peripheral auditory

processing. Different ways to model peripheral functions are employed in order to include additional aspects of auditory perception.

After initial computation of short-time features, the robustness of the feature vector sequence can be increased by certain post-processing operations which typically consider longer time intervals and many frames. Such an approach can easily boost certain *modulation frequencies*. Modulation frequency selectivity is another, high-level (neural) perceptual function whose accurate modeling may improve system performance. Finally, some perceptual approaches aim to integrate short- and long- or medium-term processing.

4.4.1 Robust spectrum analysis

The basic method of spectrum analysis, which makes minimal assumptions about the signal, is the periodogram, which is implemented by dividing the squared magnitude of the DFT by the number of points in the DFT (Eq. 2.8). It works well with a sufficient amount of clean observations of the data, but runs into problems with short analysis frames containing noisy observations [53]. In one study, simply differentiating the power spectrum with respect to frequency in a MFCC-like framework improved the noise robustness in speech recognition experiments [24]. Other *non-parametric* spectrum estimation techniques [53], such as the Thomson [128] and multitaper methods have led to improved noise robustness in speaker verification [72]. Replacing the MFCC filterbank with another *speech-signal-based* triangular filterbank, optimized on large speech corpora so that filter bandwidth is inversely proportional to area in the logarithmic average short-time speech spectrum, has also shown promise [95].

A *parametric* alternative to DFT is to assume that the signal follows an autoregressive (AR) model and to use LP to generate an estimate of the spectrum envelope (Section 2.2.2). In different studies, automatic speech and speaker recognition performance has been found to improve in mismatched noisy conditions simply by substituting LP for DFT as the spectrum analysis method, e.g. [28, 101]. *Time-weighted* variants of LP, such as weighted linear prediction (WLP), have been observed to lead to further potential performance gains [101, 103]. Because of the promise shown by time-weighted linear predictive methods in feature extraction for machine learning applications, this thesis focuses on the time-weighted, parametric approach in the context of robust spectrum analysis.

WLP was originally introduced by Ma et al. in [81] and applied in the context of formant analysis. It is a generalization of LP in which the error energy to

be minimized,

$$E_{\mathrm{WLP}} = \sum_n \left(s_n - \sum_{k=1}^{p} a_k s_{n-k} \right)^2 W_n, \tag{4.10}$$

involves a temporal weighting of the squared prediction error by the weighting function W_n. While Ma et al. investigated both discrete- and continuous-valued weighting, in subsequent applications, this weighting function has typically been chosen as the short-time energy (STE) of the p samples prior to the predicted sample (the p samples on which linear prediction with order p is based):

$$W_n = \sum_{i=1}^{p} s_{n-i}^2. \tag{4.11}$$

This weighting places emphasis on high-energy regions, as they can be assumed to be less affected by stationary additive noise; insofar as the SPL of the noise stays the same, larger values of the STE must reflect higher SPLs of the source signal (speech) and, therefore, higher local SNRs. Naturally, the STE will also be affected by transients and non-stationarity of the noise. With sufficiently low overall SNR, the within-frame dynamic variation of non-stationary noise will overwhelm that of speech and render STE weighting ineffective.

Similarly to conventional LP (Section 2.2.2), setting the partial derivatives of Eq. 4.10 with respect to each a_k to zero yields the normal equations

$$\sum_{k=1}^{p} a_k \sum_n W_n s_{n-k} s_{n-i} = \sum_n W_n s_n s_{n-i}, \quad 1 \leq i \leq p, \tag{4.12}$$

whose solution gives the WLP model. When W_n is a nonzero constant, it cancels out from both sides of the equations, and WLP reduces to conventional LP. With the notation $\boldsymbol{s}_n = (s_{n-1} \dots s_{n-p})'$ and $\boldsymbol{a} = (a_1 \dots a_p)'$, the normal equations can be expressed in matrix form as

$$\left(\sum_n W_n \boldsymbol{s}_n \boldsymbol{s}_n' \right) \boldsymbol{a} = \sum_n W_n s_n \boldsymbol{s}_n. \tag{4.13}$$

Both sides of Eq. 4.13 involve weighted sums of autocorrelation "snapshots"; instantaneous matrices $\boldsymbol{s}_n \boldsymbol{s}_n'$ and vectors $s_n \boldsymbol{s}_n$ are summed on the left-hand side and the right-hand side, respectively. Incidentally, the commonly used STE weighting function $W_n = \sum_{i=1}^{p} s_{n-i}^2$ is the squared Euclidean norm of vector $\boldsymbol{s}_n$ and the Frobenius norm (as well as the trace) of matrix $\boldsymbol{s}_n \boldsymbol{s}_n'$.

Publications I, II and IV to VI of this thesis specifically investigate time-weighted linear predictive methods, all of which are relatives of the basic

WLP method described above. In general, they use more elaborate weighting schemes than the straightforward scheme given by Eqs. 4.10 and 4.11.

Several other parametric spectrum analysis methods have been published with a stated aim of robustness with respect to some effect, e.g., [37, 76, 88].

Minimum variance spectrum estimation, originally proposed by Capon in 1969 [22, 53] and applied to speech processing by Murthi and Rao [88] in the constrained variant of *minimum variance distortionless response* (MVDR), has found application in robust feature extraction for ASR [31, 32, 134, 138]. In general, MVDR produces an all-pole model which has a particularly smooth envelope spectrum. In the first ASR studies [31, 32, 134], it was applied as a replacement of the FFT spectrum estimate in MFCC computation in order to improve additive-noise robustness. Publication I compares MVDR in this kind of framework against another spectrally smooth all-pole method which is a descendant of WLP.

The linear-predictive approach proposed by Lee [76] and the *discrete all-pole modeling* proposed by El-Jaroudi and Makhoul [37] have been used as robust formant estimation methods. In Publication V, they are among the methods evaluated in F0-robust formant analysis together with variants of WLP.

4.4.2 Improved perceptual models

The human auditory system can be considered robust in terms of dealing with degraded auditory input, at least in comparison to basic recognition techniques such as feature extraction with MFCCs and delta coefficients. These basic techniques do already take into account some perceptual knowledge in the feature representation. Nevertheless, many methods have achieved increased robustness, especially in the field of ASR, using more careful emulation of the functionality of the auditory system [57].

In addition to MFCCs, *perceptual linear prediction* (PLP) [56] is another popular auditory spectral feature extraction technique. The technique first weights the auditory filterbank outputs with an equal loudness curve. Then the intensity-loudness compression is simulated by taking an approximate cubic root (power 0.33) of the weighted filter outputs. The resulting auditory power spectrum is transformed into an autocorrelation sequence by determining the IDFT and, finally, an LP model is estimated using the conventional autocorrelation method. The LP-to-cepstrum conversion formulas (Eqs. 2.18-2.19) can be used to convert the PLP filter to a cepstrum. Because of the auditory frequency warping, fewer coefficients are necessary to represent the spectrum while still containing roughly the same information in a perceptual

sense. This is analogous to the mel-frequency cepstrum, which requires fewer coefficients than an ordinary cepstrum because the auditory frequency warping focuses more efficiently on the relevant information in many applications.

Perceptual MVDR-based cepstral coefficients and *Perceptual MVDR* [32, 138] are feature extraction methods that rely on combining MVDR spectrum estimation with perceptual considerations. They have shown improved robustness in automatic speech recognition.

Power-normalized cepstral coefficients [70, 71] are a newer, successful technique for improving feature extraction robustness through direct auditory cepstral analysis. In this method, feature extraction is based on a gammatone filterbank (a physiologically motivated auditory model of cochlear processing), temporal masking and noise suppression are modeled on the medium-term time scale of 50 to 120 ms and an auditorily motivated power-law nonlinearity (instead of the logarithmic compression used in, e.g., the MFCCs) is applied before the discrete cosine transform that leads to cepstral coefficients [71].

In comparison to these models, the approach chosen for feature extraction in Publications I to IV and Publication VI is to focus on robust spectrum analysis methods that do not explicitly model peripheral auditory processing. In these studies, peripheral processing is separately modeled by the conventional MFCC framework. However, the robust spectrum analysis methods under study may also be applicable in combination with some of the above mentioned perceptual feature extraction processes.

4.4.3 Feature post-processing

Popular approaches to speech feature postprocessing in order to improve the robustness of a machine learning system include short-time, frame-level methods such as *vocal tract length normalization* (VTLN) [35] and long-term methods (*modulation filtering* and *cepstral normalization*).

RASTA modulation filtering is part of the RASTA-PLP speech analysis technique proposed by Hermansky and Morgan [58]. In its original form, an IIR temporal filter with the transfer function

$$H(z) = 0.1z^4 \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}} \tag{4.14}$$

is applied to bandpass filter the time evolution (across frames) of each critical-band bin of the logarithmic auditory spectrum. With typical frame-shift intervals of about 10 ms, corresponding to frame rates of about 100 Hz, the RASTA filter focuses on the modulation frequency range that is important

both in terms of the sensitivity of hearing [7] and in terms of the modulation frequency content in speech [49]. In later studies on machine learning applications, e.g. [21], RASTA filtering has been applied as a filter which emphasizes the focus on the important, speech-specific modulation frequencies.

Let us assume that $\{S_k\}$ and $\{H_k\}$ in Eq. 2.17 are the source signal spectrum and the spectrum of a transmission channel, respectively. Furthermore, we assume that $\{H_k\}$ may change from session to session or slowly even during the same session, causing unpredictable mismatch. Then, it would be useful to subtract its effect from the composite cepstrum $\{C_j\}$ so that we would only be left with the cepstrum of the source signal. If the change of the cepstrum component IDFT $(\{\log|H_k|\})$ from frame to frame is very slow, the composite cepstrum $\{C_j\}$ can be averaged over a suitable number of frames so that the contribution of the source signal is averaged out but the slowly changing component remains. Subtraction of such a medium-term average cepstrum from $\{C_j\}$ leaves IDFT $(\{\log|S_k|\})$, an estimate of the instantaneous cepstrum of the source signal. This operation is called *cepstral mean subtraction* (CMS). To reiterate, a medium-term average cepstrum is subtracted from the cepstral feature vectors [5, 63] as

$$\hat{\boldsymbol{c}}_n = \boldsymbol{c}_n - \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{c}_{n-o-i}, \tag{4.15}$$

where $\boldsymbol{c}_n = (C_{n,1}, \ldots, C_{n,d})'$ denotes the cepstrum vector of the nth frame and o is an offset in number of frames. This has the effect of making the resulting feature vectors invariant to changes in the medium-term average that is subtracted. Thus, it serves to increase robustness against stationary or slowly changing background noise.

In addition to subtracting the medium-term cepstral mean, *cepstral mean and variance normalization* (CMVN) also normalizes the variance of each cepstral coefficient to unity within the normalization and averaging window [63].

5. Summary of the publications

Publication I: Stabilised weighted linear prediction

WLP, using the STE weighting function, is modified in a way that guarantees the resulting all-pole synthesis filters to be stable. Besides being applicable to synthesis applications, SWLP is observed to generally result in smooth spectrum models, which is often indicative of robustness. For instance, MVDR is an earlier method that also produces smooth spectrum models and has been used in order to improve robustness in recognition applications.

SWLP is evaluated both in spectrum analysis and in feature extraction for isolated word recognition in terms of robustness against increasing additive noise corruption. In comparison to the conventional spectrum estimation methods FFT and LP as well as MVDR, SWLP shows the best overall robustness performance.

Publication II: Temporally weighted linear prediction features for tackling additive noise in speaker verification

In this study, WLP and SWLP are applied as spectrum analysis methods in feature extraction for text-independent speaker verification.

It is found that the weighted all-pole methods WLP and especially SWLP improve upon conventional FFT and LP spectrum analyses in terms of additive-noise robustness when each of these methods in turn is used as the initial spectrum analysis method in MFCC feature extraction. WLP and SWLP also show minor improvement over FFT in the clean condition, which may be indicative of robustness with respect to speaker or telephone channel effects. Speech enhancement by spectral subtraction as a preprocessing step is found to lead to large performance improvement in the noise-corrupted cases.

However, the performance gain of SWLP, in particular, over the conventional methods is preserved also in the cases with speech enhancement preprocessing.

Publication III: Detection of shouted speech in noise: human and machine

A system is developed for detecting shouted speech events in the presence of ambient acoustic noise by focusing on robust short-time spectral feature extraction. The detection performance of the system is evaluated against that of human listeners.

Two things in particular are found to improve additive-noise robustness: (1) replacing conventional FFT spectrum analysis in MFCC feature extraction by a method which multiplies a linear predictive envelope with a cepstrally separated spectral fine structure that represents the vocal tract excitation and (2) using a longer-than-normal MFCC feature vector without delta coefficients. Furthermore, in the spectral multiplication in (1), adaptive weighting of the excitation spectrum based on the spectral flatness of the envelope, as a measure of the noisiness of the frame, is used. Figure 5.1 illustrates the spectral multiplication. The automatic detection system outperforms human listeners at moderate SNRs of multitalker noise and matches human performance at low SNRs.

Publication IV: Extended weighted linear prediction using the autocorrelation snapshot - a robust speech analysis method and its application to recognition of vocal emotions

The thus far most general formulation of time-weighted linear prediction, whose focus on the information contained within the analysis frames can be adjusted with few constraints, is presented. Implementations of the method are applied to improving the additive-noise robustness of spectrum models in feature extraction for speech emotion recognition.

It is found that two new information weighting schemes, made possible by the new general formulation, both improve the robustness of the emotion recognition system in cases with additive-noise mismatch. Figure 5.2 shows spectra over one utterance obtained with conventional LP, WLP and extended weighted linear prediction using a partial-weight approach (XLP-P) and the two new weighting schemes (XLP-S1 and XLP-S2).

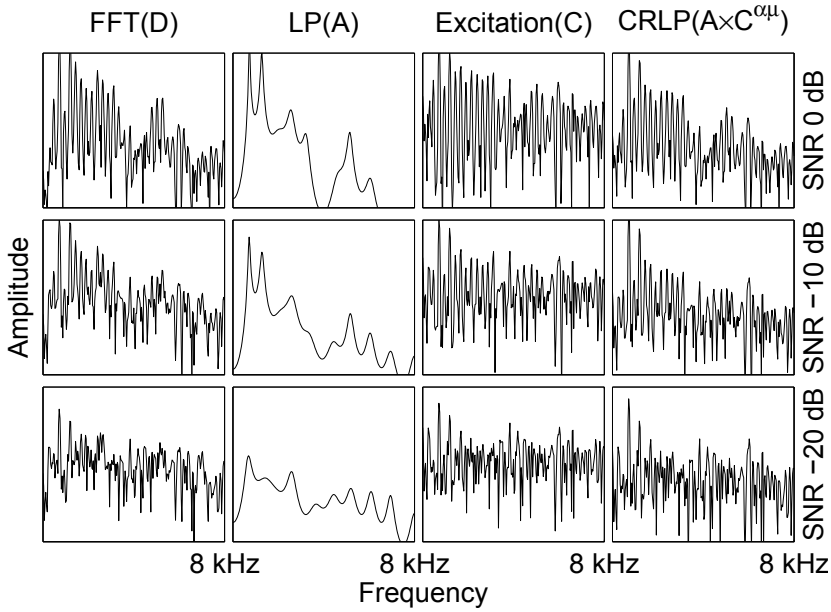

Figure 5.1. Example spectra based on a shouted vowel frame by a male speaker. The rows correspond, from top to bottom, to SNR levels 0 dB, −10 dB, and −20 dB with factory noise corruption. The columns correspond to different types of spectra.

Publication V: Formant frequency estimation of high-pitched vowels using weighted linear prediction

A weighting function for WLP is designed to down-weight the contribution of voiced speech excitation in the vicinity of glottal closure instants (GCIs). In this study, the GCIs are obtained in the manner of an "oracle", both by using synthetic speech material, where the GCI locations are well known, and by using natural speech material recorded simultaneously with an electroglottography signal.

The results indicate that the formant bias, which particularly affects linear predictive models when the fundamental frequency F0 is high, is greatly reduced by down-weighting the GCIs. With the help of the oracle knowledge, the proposed approach shows clear improvement not only upon conventional LP, but also upon previous linear predictive methods proposed for robust formant estimation. The results thus support the use of the WLP approach for F0-robust formant estimation. In Publication VI, the same method is used to improve glottal inverse filtering with GCI locations determined automatically from the speech signal [1].

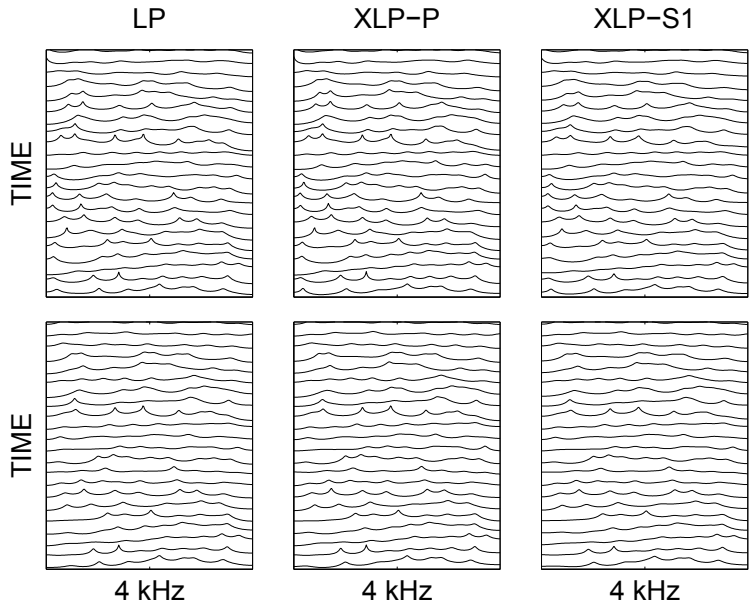

Figure 5.2. Example spectra of LP, XLP-P and XLP-S1 over one utterance of the anger emotion category. Upper panels: clean speech. Lower panels: the same utterance with noise corruption by factory noise at SNR 0 dB.

Publication VI: Mixture linear prediction in speaker verification under vocal effort mismatch

A stochastic approach to weighted linear predictive spectrum modeling of a signal frame, based on a mixture autoregressive model whose one state (component) is considered as the *target*, is presented. The method has the ability to focus on different aspects of the signal according to initialization of the spectral and amplitude characteristics of target and *non-target* states prior to iterative parameter estimation using the EM algorithm. It is shown that the proposed method can be viewed as a form of WLP weighted using state posterior probabilities of the mixture model. Motivated by the results of Publication V, a special case of the proposed general method is developed to tackle the issue of formant bias due to the fundamental frequency F0 by avoiding GCIs in the temporal domain. In speech, changes in F0 can be induced, e.g., by changes in vocal effort (e.g., alternating between normal and shouted voice). The application under study is text-independent speaker verification with mismatch induced by variation of vocal effort, and the proposed method is applied in order to improve the robustness.

The proposed method is found in the experiments to improve the vocal-effort robustness of two speaker verification systems by avoiding the GCIs in

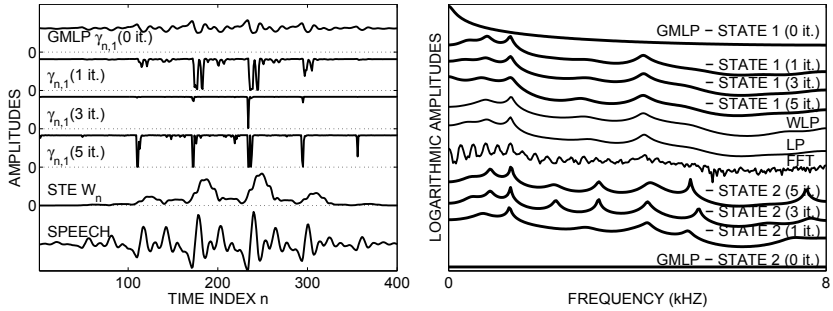

Figure 5.3. Left: A Hamming-windowed speech frame (vowel from a female speaker sampled at 16 kHz) with different weighting functions. STE is the weighting scheme generally used with WLP. GMLP weights, which tend to avoid GCIs, result from iterative EM re-estimation based on the initial autoregression templates. Right: The corresponding spectra of FFT, LP, WLP, and GMLP ($p = 20$), including the initial spectra of the GMLP states.

probabilistic weighting of the signal information. Figure 5.3 illustrates the evolution of the spectra associated with two mixture components, and the corresponding probabilistic weighting function, during the course of iterative EM model parameter estimation.

Publication VII: Multi-scale modulation filtering in automatic detection of emotions in telephone speech

Autoregressive filters on several time scales are used to predict short-time acoustic features of speech. Such prediction is used as filtering to emphasize modulation frequencies specific to a target emotion class in detecting activation and valence in telephone speech.

The results on the Berlin database with simulated far-end noise corruption and narrowband telephone channel show that detection robustness is improved by the proposed filtering method. When combined with automatic training data selection, also based on the filtering method, valence detection is also improved in the clean test condition, suggesting that the method can also tackle speaker-related mismatch. Figure 5.4 shows the effect of the filtering on auditory spectra transformed back from mel-frequency cepstral coefficients.

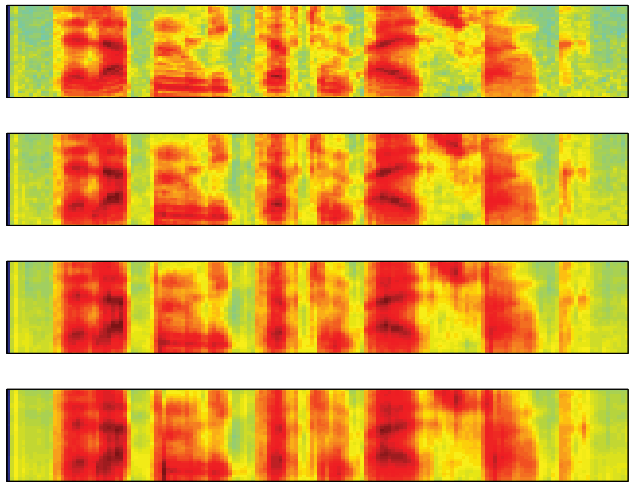

Figure 5.4. Top panel: mel-scale spectrogram, with 40 bins, transformed back from MFCCs for a neutral telephone utterance (original label 03a01Nc) corrupted by car interior noise (SNR 0 dB). Lower panels: mel-scale spectrograms for the same utterance after filtering the original MFCCs with multi-scale autoregressive predictors for classes "anger", "neutral" and "happiness".

Publication VIII: Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits

A feature-intensive approach to paralinguistic speech classification is investigated from the point of view of feature selection. Several new feature selection methods are evaluated individually and in different combinations in classification problems in which a large and comprehensive set of long-term (utterance-level) features is available, but the amount of training data is very small for the feature space dimensionality.

The newly proposed *supervised* feature selection methods–random subset feature selection (RSFS), feature selection as a set-covering problem of supervised classifications (SSCP) and statistical-dependency feature selection (SD)–each provide positive results in terms of (1) similar or improved classification accuracy (using a kNN classifier) as compared to the full feature set; (2) similar or improved classification accuracy in comparison to established feature selection methods, sequential forward selection (SFS) and minimum-redundancy-maximum-relevance (MRMR); and (3) clearly reduced computational load with respect to MRMR and SFS. In addition to the supervised methods, completely or partially *unsupervised methods* are found to be able to improve feature selection results when included in various combinations of feature selection methods.

6. Conclusions

This thesis consists of studies on robust speech processing methods. They tackle spectral feature extraction using all-pole models (Publications I to VI), modulation filtering that emphasizes class-specific long-term dynamics of feature vector sequences (Publication VII) and feature selection for long-term (utterance-level) features (Publication VIII). A common theme in all studies was to find ways to better emphasize and uncover information that may benefit analysis and machine learning systems in speech processing. The applications under study fall into the broad categories of ASR (Publication I), speaker recognition (Publications II and VI), speech event detection (Publication III), speech paralinguistics (Publications IV, VII and VIII) and formant analysis (Publication V).

The modeling of human auditory perception has often been found beneficial in improving the robustness of speech analysis systems subject to variable operating conditions, such as additive noise, channel variation and speaker-related variation in speaking style. The eight studies in this thesis can be divided into those that concern short-term (Publications I to VI) and long-term (Publications VII and VIII) modeling of the acoustic properties of speech. In the physiological sense, short-term and long-term modeling can be roughly associated with the auditory periphery and the neural auditory pathway, respectively. In the perceptual or psychoacoustical sense, short-term methods model timbre, pitch and loudness, while long-term methods model subjective duration or rhythm.

In general, the relative importance of short-term and long-term modeling is determined by the application at hand. In paralinguistic analysis, for example, long-term modeling is typically important in order to distinguish between different types of speech which may have similar short-term characteristics but differ in the long-term modulation frequency content. This is demonstrated by the results of Publications VII and VIII. These studies used different methods

of emphasizing long-term information in order to achieve improved recognition accuracy. In Publication VII, autoregressive modulation filtering of short-term features on multiple time scales improved the detection of affective dimensions over the baseline, which used only timbral features. Improvement was obtained both with clean speech (speaker-related mismatch only) and in the presence of additive-noise mismatch. This filtering approach, which emphasizes the modulation dynamics of its trained target class, is generic and modular and thus potentially applicable in various problems as a feature postprocessing step. In Publication VIII, starting from a large and comprehensive set of different long-term features that parametrized different acoustic effects over utterances, combinations of novel feature selection methods were able to find feature subsets that better focused on the target classes despite speaker-related variability. A combination of different kinds of supervised and unsupervised feature selection criteria was found to be a promising approach. The proposed feature selection methods achieved or exceeded the performance level of the earlier feature selection approaches evaluated, but with a reduced computational cost. Analysis of the types of long-term features uncovered by several different feature selection criteria showed that different features performed best in the seven different speaker trait recognition tasks.

On the short term, robust spectrum analysis methods were studied. A stabilized version of WLP (Publication I) improved the robustness of isolated word recognition (Publication I) and text-independent speaker verification (Publication II) with respect to additive noise corruption. A generalized formulation of time-weighted linear prediction was presented and applied to improving the additive-noise robustness of speech emotion recognition (Publication IV). In Publication III, the spectrum envelope obtained using a linear predictive method and the spectral fine structure obtained using cepstral analysis were combined by multiplication in order to improve the robustness of a shout-event detection system in the presence of ambient noise. Publication V applied "oracle" weighting functions in WLP in order to exclude the contribution of the voiced excitation of speech and to reduce the biasing of formant estimates. Publication VI presented a stochastic mixture decomposition approach to linear predictive modeling and applied it to robust feature extraction in speaker verification under vocal effort mismatch.

Many problems of automatic speech analysis are complex and subject to large variability in terms of conditions such as noise, channel and speaking style. A large number of features is typically needed to represent all of the short-term and long-term acoustic effects that can possibly be relevant for a

given analysis problem. This easily leads to problems of insufficient training data brought about by high dimensionality. Instead of tackling this issue with more complex and/or specialized pattern recognition and machine learning models, this thesis set out to develop feature extraction methods that can be customized according to the analysis problem to better focus on important aspects while limiting the influence of misleading or irrelevant information. The results of Publications I to VIII indicate that the robustness of speech pattern recognition and speech analysis can indeed be significantly improved by the approach of making feature extraction more specific to the analysis problem at hand–an idea that is complementary to the currently prevalent black-box/brute-force trend in the sense that it may be able to further improve the performance of those systems. Considering that the methods proposed are, by their nature, nevertheless rather generic and mostly well customizable, they are believed to have potential for future applicability in various feature extraction and analysis problems in speech, audio and signal processing applications.

Bibliography

[1] M. Airaksinen, T. Raitio, B. Story, and P. Alku. Quasi closed phase glottal inverse filtering analysis with weighted linear prediction. *IEEE Transactions on Audio, Speech and Language Processing*, 22(3):596–607, March 2014.

[2] M. B. Al-Daoud and S. A. Roberts. New methods for the initialisation of clusters. *Pattern Recognition Letters*, 17(5):451–455, May 1996.

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

[4] E. Alpaydin. *Introduction to Machine Learning*. The MIT Press, second edition, 2010.

[5] B. S. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustical Society of America*, 55(6):1304–1312, June 1974.

[6] M. Athineos and D. P. W. Ellis. Frequency-domain linear prediction for temporal features. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'03)*, pages 261–266, Saint Thomas, Virgin Islands, USA, November 30–December 3 2003.

[7] S. P. Bacon and N. F. Viemeister. Temporal modulation transfer functions in normal-hearing and hearing-impaired listeners. *Audiology*, 24:117–134, 1985.

[8] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens. Automatic speech recognition and speech variability: a review. *Speech Communication*, 49(10–11):763–786, October–November 2007.

[9] J. A. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Technical report, University of California at Berkeley, April 1998.

[10] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer Science+Business Media, LLC, 2006.

[11] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2):113–120, April 1979.

[12] D. Bone, M. P. Black, M. Li, A. Metallinou, S. Lee, and S. S. Narayanan. Intoxicated speech detection by fusion of speaker normalized hierarchical features

and GMM supervectors. In *Proc. Interspeech 2011*, pages 3217–3220, Florence, Italy, August 28–31 2011.

[13] L. Bottou and Y. Bengio. Convergence properties of the k-means algorithms. In *Advances in Neural Information Processing Systems 7*. MIT Press, 1995.

[14] P. S. Bradley and U. M. Fayyad. Refining initial points for k-means clustering. In *Proc. 15th International Conference on Machine Learning (ICML '98)*, pages 91–99, Madison, Wisconsin, USA, July 24–27 1998.

[15] A. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. The MIT Press, 1990.

[16] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, October 2001.

[17] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth International, 1984.

[18] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, June 1998.

[19] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. A database of German emotional speech. In *Proc. Interspeech 2005*, pages 1517–1520, Lisbon, Portugal, September 4–8 2005.

[20] C. Busso and T. Rahman. Unveiling the acoustic properties that describe the valence dimension. In *Proc. Interspeech 2012*, Portland, Oregon, USA, September 9–13 2012.

[21] W. M. Campbell, D. E. Sturim, and D. A. Reynolds. Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*, 13(5):308–311, May 2006.

[22] J. Capon. High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE*, 57(8):1408–1418, August 1969.

[23] O. Cappé. Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. *IEEE Transactions on Speech and Audio Processing*, 2(2):345–349, April 1994.

[24] J. Chen, K. Paliwal, and S. Nakamura. Cepstrum derived from differentiated power spectrum for robust speech recognition. *Speech Communication*, 41(2–3):469–484, October 2003.

[25] T.-S. Chi, L.-Y. Yeh, and C.-C. Hsu. Robust emotion recognition by spectro-temporal modulation statistic features. *Journal of Ambient Intelligence and Humanized Computing*, 3(1):47–60, March 2012.

[26] M. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, 34(3):267–285, June 2001.

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

[28] F. de Wet, B. Cranen, J. de Veth, and L. Boves. A comparison of LPC and FFT-based acoustic features for noise robust ASR. In *Proc. Eurospeech 2001*, Aalborg, Denmark, September 3–7 2001.

[29] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, May 2011.

[30] A. P. Dempster, N. M. Laird, and D. B.Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, 39:1–38, 1977.

[31] S. Dharanipragada and B. D. Rao. MVDR based feature extraction for robust speech recognition. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2001)*, pages 309–312, Salt Lake City, Utah, USA, May 7–11 2001.

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

[33] T. Drugman, J. Kane, and C. Gobl. Resonator-based creaky voice detection. In *Proc. Interspeech 2012*, Portland, Oregon, USA, September 9–13 2012.

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

[35] E. Eide and H. Gish. A parametric approach to vocal tract length normalization. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '96)*, pages 346–348, Atlanta, Georgia, USA, May 7–10 1996.

[36] M. El Ayadi, M. S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, March 2011.

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

[38] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(6):1109–1121, December 1984.

[39] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(2):443–445, April 1985.

[40] M. Erden and L. M. Arslan. Automatic detection of anger in human-human call center dialogs. In *Proc. Interspeech 2011*, pages 81–84, Florence, Italy, August 28–31 2011.

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

[42] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, September 1936.

[43] H. Fletcher. *Speech and Hearing in Communication*. New York: Krieger, 1953.

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

[45] S. Ganapathy and H. Hermansky. Temporal resolution analysis in frequency domain linear prediction. *Journal of the Acoustical Society of America*, 132(5):EL436–EL442, November 2012.

[46] A. Ghias, J. Logan, D. Chamberlin, and B. C. Smith. Query by humming - musical information retrieval in an audio database. In *Proc. Third ACM International Conference on Multimedia (ACM Multimedia '95)*, pages 231–236, San Francisco, USA, November 5–9 1995.

[47] K. W. Godin and J. H. L. Hansen. Analysis and perception of speech under physical task stress. In *Proc. Interspeech 2008*, pages 1674–1677, Brisbane, Australia, September 22–26 2008.

[48] A. H. Gray and J. D. Markel. Distance measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(5):380–391, October 1976.

[49] S. Greenberg. On the origins of speech intelligibility in the real world. In *Proc. ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, pages 23–32, Pont-a-Mousson, France, April 17–18 1997.

[50] J. H. L. Hansen and M. A. Clements. Constrained iterative speech enhancement with application to speech recognition. *IEEE Transactions on Signal Processing*, 39(4):795–805, April 1991.

[51] J. H. L. Hansen and S. A. Patil. Speech under stress: analysis, modeling and recognition. In *Speaker Classification I, Lecture Notes in Computer Science*, pages 108–137. Springer-Verlag, 2007.

[52] A. Härmä, M. F. McKinney, and J. Skowronek. Automatic surveillance of the acoustic activity in our living environment. In *Proc. IEEE International Conference on Multimedia and Expo (ICME 2005)*, Amsterdam, The Netherlands, July 6–8 2005.

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

[54] S. Haykin. *Adaptive Filter Theory*. Prentice-Hall, third edition, 1996.

[55] J. He, M. Lan, C.-L. Tan, S.-Y. Sung, and H.-B. Low. Initialization of cluster refinement algorithms: A review and comparative study. In *Proc. IEEE International Joint Conference on Neural Networks (IJCNN 2004)*, pages 297–302, Budapest, Hungary, July 25–29 2004.

[56] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, April 1990.

[57] H. Hermansky, J. R. Cohen, and R. M. Stern. Perceptual properties of current speech recognition technology. *Proceedings of the IEEE*, 101(9):1968–1985, September 2013.

[58] H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, October 1994.

[59] W. Hess. *Pitch Determination of Speech Signals*. Springer-Verlag, 1983.

[60] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, November 2012.

[61] T. K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, August 1998.

[62] Y. Hu and P. C. Loizou. Speech enhancement based on wavelet thresholding the multitaper spectrum. *IEEE Transactions on Speech and Audio Processing*, 12(1):59–67, January 2004.

[63] X. Huang, A. Acero, and H.-W. Hon. *Spoken Language Processing*. Prentice Hall PTR, 2001.

[64] F. Itakura. Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(1):67–72, February 1975.

[65] F. Jelinek. *Statistical Methods for Speech Recognition*. The MIT Press, 1997.

[66] W. Jesteadt, S. P. Bacon, and J. R. Lehman. Forward masking as a function of frequency, masker level, and signal delay. *Journal of the Acoustical Society of America*, 71(4):950–962, April 1982.

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

[68] I. Katsavounidis, C.-C. J. Kuo, and Z. Zhang. A new initialization technique for generalized Lloyd iteration. *IEEE Signal Processing Letters*, 1(10):144–146, October 1994.

[69] S. Keronen, J. Pohjalainen, P. Alku, and M. Kurimo. Noise robust feature extraction based on extended weighted linear prediction in LVCSR. In *Proc. Interspeech 2011*, pages 1265–1268, Florence, Italy, August 28–31 2011.

[70] C. Kim and R. M. Stern. Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction. In *Proc. Interspeech 2009*, pages 28–31, Brighton, UK, September 6–10 2009.

[71] C. Kim and R. M. Stern. Power-normalized cepstral coefficients (PNCC) for robust speech recognition. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, pages 4101–4104, Kyoto, Japan, March 25–30 2012.

[72] T. Kinnunen, R. Saeidi, F. Sedlak, K. A. Lee, J. Sandberg, M. Hansson-Sandsten, and H. Li. Low-variance multitaper MFCC features: a case study in robust speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, 20(7):1990–2001, September 2012.

[73] B. Kollmeier and R. Koch. Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction. *Journal of the Acoustical Society of America*, 95(3):1593–1602, March 1994.

[74] J. Krajewski, A. Batliner, and M. Golz. Acoustic sleepiness detection: Framework and validation of a speech-adapted pattern recognition approach. *Behavior Research Methods*, 41(3):795–804, August 2009.

[75] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan. Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, 53(9–10):1162–1171, November–December 2011.

[76] C.-H. Lee. On robust linear prediction of speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(5):642–650, May 1988.

[77] M. Li, K. J. Han, and S. Narayanan. Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. *Computer Speech and Language*, 27(1):151–167, January 2013.

[78] J. Lim and A. V. Oppenheim. All-pole modeling of degraded speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(3):197–210, June 1978.

[79] P. C. Loizou. *Speech Enhancement: Theory and Practice*. CRC Press, 2007.

[80] L.-S. A. Low, M. C. Maddage, M. Lech, and L. B. Sheeber. Detection of clinical depression in adolescents' speech during family interactions. *IEEE Transactions on Biomedical Engineering*, 58(3):574–586, March 2011.

[81] C. Ma, Y. Kamp, and L.F. Willems. Robust signal selection for linear prediction analysis of voiced speech. *Speech Communication*, 12(2):69–81, March 1993.

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

[83] T. Marill and D. Green. On the effectiveness of receptors in recognition systems. *IEEE Transactions on Information Theory*, 9(1):11–17, January 1963.

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

[85] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance. In *Proc. Eurospeech 1997*, Rhodes, Greece, September 22–25 1997.

[86] B. C. J. Moore, editor. *Hearing*. Academic Press, 1995.

[87] D. P. Morgan and C. L. Scofield. *Neural Networks and Speech Processing*. Kluwer Academic Publishers, 1991.

[88] M. N. Murthi and B. D. Rao. All-pole model parameter estimation for voiced speech. In *Proc. 1997 IEEE Workshop on Speech Coding for Telecommunications*, pages 17–18, Pocono Manor, Pennsylvania, USA, September 7–10 1997.

[89] D. Neiberg, K. Elenius, and K. Laskowski. Emotion recognition in spontaneous speech using GMMs. In *Proc. Interspeech 2006*, pages 809–812, Pittsburgh, Pennsylvania, USA, September 17–21 2006.

[90] N. Obin. Cries and whispers–classification of vocal effort in expressive speech. In *Proc. Interspeech 2012*, Portland, Oregon, USA, September 9–13 2012.

[91] A. V. Oppenheim and R. W. Schafer. *Digital Signal Processing*. Prentice-Hall, 1975.

[92] D. O'Shaughnessy. *Speech Communications: Human and Machine*. IEEE Press, second edition, 2000.

[93] K. Paliwal and L. Alsteris. Usefulness of phase spectrum in human speech perception. In *Proc. Eurospeech 2003*, pages 2117–2120, Geneva, Switzerland, September 1–4 2003.

[94] K. Paliwal, B. Schwerin, and K. Wójcicki. Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator. *Speech Communication*, 54(2):282–305, February 2012.

[95] K. Paliwal, B. Shannon, J. Lyons, and K. Wójcicki. Speech-signal-based frequency warping. *IEEE Signal Processing Letters*, 16(4):319–322, April 2009.

[96] K. Paliwal, K. Wójcicki, and B. Schwerin. Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. *Speech Communication*, 52(5):450–475, May 2010.

[97] S. A. Patil and J. H. L. Hansen. Detection of speech under physical stress: model development, sensor selection, and feature fusion. In *Proc. Interspeech 2008*, pages 817–820, Brisbane, Australia, September 22–26 2008.

[98] J. M. Peña, J. A. Lozano, and P. Larrañaga. An empirical comparison of four initialization methods for the k-means algorithm. *Pattern Recognition Letters*, 20(10):1027–1040, October 1999.

[99] J. Pohjalainen and P. Alku. Automatic detection of anger in telephone speech with robust autoregressive modulation filtering. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, pages 7537–7541, Vancouver, Canada, May 26–31 2013.

[100] J. Pohjalainen, S. Kadioglu, and O. Räsänen. Feature selection for speaker traits. In *Proc. Interspeech 2012*, Portland, Oregon, USA, September 9–13 2012.

[101] J. Pohjalainen, H. Kallasjoki, K.J. Palomäki, M. Kurimo, and P. Alku. Weighted linear prediction for speech analysis in noisy conditions. In *Proc. Interspeech 2009*, pages 1315–1318, Brighton, UK, September 6–10 2009.

[102] J. Pohjalainen, T. Raitio, H. Pulakka, and P. Alku. Automatic detection of high vocal effort in telephone speech. In *Proc. Interspeech 2012*, Portland, Oregon, USA, September 9–13 2012.

[103] J. Pohjalainen, R. Saeidi, T. Kinnunen, and P. Alku. Extended weighted linear prediction (XLP) analysis of speech and its application to speaker verification in adverse conditions. In *Proc. Interspeech 2010*, pages 1477–1480, Makuhari, Japan, September 26–30 2010.

[104] G. E. Poliner, D. P. W. Ellis, A. F. Ehmann, E. Gomez, S. Streich, and B. Ong. Melody transcription from musical audio: Approaches and evaluation. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1247–1256, May 2007.

[105] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 1992.

[106] P. Pudil, J. Novovicová, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15(11):1119–1125, November 1994.

[107] J. Pylkkönen and M. Kurimo. Improving discriminative training for robust acoustic models in large vocabulary continuous speech recognition. In *Proc. Interspeech 2012*, Portland, Oregon, USA, September 9–13 2012.

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

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

[110] R. Radhakrishnan, A. Divakaran, and P. Smaragdis. Audio analysis for surveillance applications. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'05)*, pages 158–161, New Paltz, New York, USA, October 16–19 2005.

[111] O. Räsänen and J. Pohjalainen. Random subset feature selection in automatic recognition of developmental disorders, affective states, and level of conflict from speech. In *Proc. Interspeech 2013*, pages 210–214, Lyon, France, August 25–29 2013.

[112] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1–3):19–41, January 2000.

[113] D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83, January 1995.

[114] T. D. Rossing, F. R. Moore, and P. A. Wheeler. *The Science of Sound*. Addison Wesley, third edition, 2002.

[115] J.-L. Rouas, J. Louradour, and S. Ambellouis. Audio events detection in public transport vehicle. In *Proc. IEEE Intelligent Transportation Systems Conference (ITSC 2006)*, pages 733–738, Toronto, Canada, September 17–20 2006.

[116] P. Scalart and J. Filho. Speech enhancement based on a priori signal to noise estimation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '96)*, pages 629–632, Atlanta, Georgia, USA, May 7–10 1996.

[117] S. M. Schimmel, L. E. Atlas, and K. Nie. Feasibility of single channel speaker separation based on modulation frequency analysis. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, pages 605–608, Honolulu, Hawaii, USA, April 15–20 2007.

[118] B. Schuller, D. Arsic, F. Wallhoff, and G. Rigoll. Emotion recognition in the noise applying large acoustic feature sets. In *Proc. Speech Prosody*, Dresden, Germany, May 2–5 2006.

[119] B. Schuller and A. Batliner. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. John Wiley and Sons Inc., 2013.

[120] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan. Paralinguistics in speech and language - state-of-the-art and the challenge. *Computer Speech and Language*, 27(1):4–39, January 2013.

[121] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang. The Interspeech 2014 Computational Paralinguistics Challenge: Cognitive & physical load. In *Proc. Interspeech 2014*, pages 427–431, Singapore, September 14–18 2014.

[122] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss. The Interspeech 2012 Speaker Trait Challenge. In *Proc. Interspeech 2012*, Portland, Oregon, USA, September 9–13 2012.

[123] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, E. Salamin, A. Polychroniou, F. Valente, and S. Kim. The Interspeech 2013 Computational Paralinguistics Challenge: Social signals, conflict, emotion, autism. In *Proc. Interspeech 2013*, pages 148–152, Lyon, France, August 25–29 2013.

[124] E. Shriberg, M. Graciarena, H. Bratt, A. Kathol, S. S. Kajarekar, H. Jameel, C. Richey, and F. Goodman. Effects of vocal effort and speaking style on text-independent speaker verification. In *Proc. Interspeech 2008*, pages 609–612, Brisbane, Australia, September 22–26 2008.

[125] J. Sohn, N. S. Kim, and W. Sung. A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, 6(1):1–3, January 1999.

[126] E. Terhardt, G. Stoll, and M. Seewann. Algorithm for extraction of pitch and pitch salience for complex tonal signals. *Journal of the Acoustical Society of America*, 71(3):679–688, March 1982.

[127] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, second edition, 2003.

[128] D. J. Thomson. Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, 70(9):1055–1096, September 1982.

[129] M. Vacher, D. Istrate, L. Besacier, J.-F. Serignat, and E. Castelli. Sound detection and classification for medical telesurvey. In *Proc. International Conference on Biomedical Engineering*, pages 395–399, Innsbruck, Austria, February 16–18 2004.

[130] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti. Scream and gunshot detection and localization for audio-surveillance systems. In *Proc. IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS 2007)*, pages 21–26, London, UK, September 5–7 2007.

[131] L. Vidrascu and L. Devillers. Detection of real-life emotions in call centers. In *Proc. Interspeech 2005*, pages 1841–1844, Lisbon, Portugal, September 4–8 2005.

[132] A. W. Whitney. A direct method of nonparametric measurement selection. *IEEE Transactions on Computers*, C-20(9):1100–1103, September 1971.

[133] K. A. Williams and J. H. L. Hansen. Speaker height estimation combining GMM and linear regression subsystems. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, pages 7552–7556, Vancouver, Canada, May 26–31 2013.

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

[135] D. Wu, T. D. Parsons, E. Mower, and S. Narayanan. Speech emotion estimation in 3D space. In *Proc. IEEE International Conference on Multimedia and Expo (ICME 2010)*, pages 737–742, Singapore, July 19–23 2010.

[136] S. Wu, T. H. Falk, and W.-Y. Chan. Automatic speech emotion recognition using modulation spectral features. *Speech Communication*, 53(5):768–785, May–June 2011.

[137] X. Xiao, E. S. Chng, and H. Li. Normalization of the speech modulation spectra for robust speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(8):1662–1674, November 2008.

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

[139] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department, 2006.

[140] P. Zelinka and M. Sigmund. Automatic vocal effort detection for reliable speech recognition. In *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2010)*, pages 349–354, Kittilä, Finland, August 29–September 1 2010.

[141] P. Zelinka, M. Sigmund, and J. Schimmel. Impact of vocal effort variability on automatic speech recognition. *Speech Communication*, 54(6):732–742, July 2012.

[142] C. Zhang and J. H. L. Hansen. Analysis and classification of speech mode: Whispered through shouted. In *Proc. Interspeech 2007*, pages 2289–2292, Antwerp, Belgium, August 27–31 2007.

[143] E. Zwicker and H. Fastl. *Psychoacoustics, Facts and Models*. Springer-Verlag, 1990.

In addition to the linguistic message, speech carries many types of information that are related, for example, to speaker identity and attributes, speaking style and speaking situation. The use of machine learning to decipher any such information from speech is complicated by variation in the acoustic environment and the transmission channel as well as the aforementioned inherent variability of speech itself. As speech technology applications become more common, this variability constitutes a challenge. In order to tackle its effects on speech features, this thesis develops robust methods of acoustic feature extraction to be used in machine learning applications of speech technology. The specific applications under study include speech and speaker recognition, speech event detection, speech emotion recognition and speaker characterization.

Aalto-DD 203/2014

ISBN 978-952-60-6005-7 (printed)
ISBN 978-952-60-6006-4 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Electrical Engineering
Department of Signal Processing and Acoustics
www.aalto.fi

BUSINESS + ECONOMY

ART + DESIGN + ARCHITECTURE

SCIENCE + TECHNOLOGY

CROSSOVER

DOCTORAL DISSERTATIONS