Czech Technical University in Prague
Faculty of Electrical Engineering

Doctoral Thesis

March 2008

Hynek Bořil

Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Circuit Theory

Robust Speech Recognition: Analysis and Equalization of Lombard Effect in Czech Corpora

Doctoral Thesis

Hynek Bořil

Prague, March 2008

Ph.D. Programme: Electrical Engineering and Information Technology
Branch of Study: Electrical Engineering Theory

Supervisor: Doc. Ing. Petr Pollák, CSc.

Abstract

When exposed to noise, speakers will modify the way they speak in an effort to maintain intelligible communication. This process, which is referred to as Lombard effect (LE), involves a combination of both conscious and subconscious articulatory adjustment. Speech production variations due to LE can cause considerable degradation in automatic speech recognition (ASR) since they introduce a mismatch between parameters of the speech to be recognized and the ASR system's acoustic models, which are usually trained on neutral speech. The main objective of this thesis is to analyze the impact of LE on speech production and to propose methods that increase ASR system performance in LE. All presented experiments were conducted on the Czech spoken language, yet, the proposed concepts are assumed applicable to other languages.

The first part of the thesis focuses on the design and acquisition of a speech database comprised of utterances produced in neutral conditions (neutral speech), and in simulated noisy conditions (Lombard speech), and on the analysis of the speech production differences in these two speech modalities. A majority of the previous studies on the role of LE in ASR neglected the importance of the communication loop in evoking Lombard effect, and instead analyzed data from subjects who read text in noise without being provided feedback regarding whether their speech was intelligible. In this thesis, a novel setup imposes a communication factor to the Lombard recordings. An analysis of the recordings shows considerable differences between neutral and Lombard data for a number of speech production parameters. In ASR experiments, the performance of both large and small vocabulary recognizers severely degrade when switching from neutral to LE tasks.

The second part of the thesis describes the design of new methods intended to reduce the impact of LE on ASR. The methods employ *LE equalization*, *robust features*, and *model adjustments*. The goal of *LE equalization* is to transform Lombard speech tokens towards neutral before they enter the acoustic models of the ASR engine. For this purpose, a modified vocal tract length normalization and formant-driven frequency warping are designed, both significantly improving the recognition performance under LE. In addition, a commercial voice conversion framework is evaluated and found to be partially effective for LE-equalization. A set of *robust features* are proposed in a data-driven design. Filter banks better reflecting the distribution of linguistic content in frequency are constructed and used as replacements for mel and Bark filter banks in MFCC (mel frequency cepstral coefficients) and PLP (perceptual linear prediction) front-ends. When employed in a recognition system on LE data, the novel features considerably outperform standard MFCC and PLP front-ends as well as state-of-the-art MR–RASTA (multi-resolution relative spectra) and Expolog front-ends. In the domain of *model adjustments*, an independently furnished acoustic model adaptation, which transforms neutral models towards Lombard speech characteristics, is shown to provide a substantial performance improvement on LE speech data. Finally, a two-stage recognition system (TSR) utilizing neutral/LE classification and style-specific acoustic modeling is proposed. Compared to multi-stage systems presented in other studies, TSR requires only neutral samples for training the style-specific models. On the mixture of neutral and Lombard utterances, TSR also significantly outperforms discrete style-specific recognizers. These contributions serve to advance both knowledge and algorithm development for speech recognition in Lombard effect.

Abstrakt

Vystaveni hlučnému prostředí, mluvčí mění způsob, jakým mluví, ve snaze dosáhnout srozumitelné komunikace. Tento proces, nazývaný Lombardův efekt (LE), představuje kombinaci vědomých a podvědomých změn artikulace. Změny řečové produkce vyvolané LE mohou způsobit značné zhoršení přesnosti automatického rozpoznávání řeči (ASR) v důsledku rozdílu mezi parametry zpracovávané promluvy a akustickými modely ASR systému, obvykle trénovanými na neutrálních promluvách. Hlavním cílem této disertační práce je analýza dopadu LE na parametry řečové produkce a návrh metod zvyšujících odolnost ASR systémů vůči LE. Všechny presentované experimenty byly prováděny na českých promluvách, nicméně lze očekávat, že předkládané koncepce budou použitelné i v jiných jazycích.

První část disertace se zabývá návrhem a sběrem databáze obsahující promluvy produkované v neutrálních podmínkách (neutrální řeč) a simulovaných hlučných podmínkách (Lombardovu řeč) a analýzou změn řečové produkce v těchto dvou modalitách. Většina předchozích studií věnovaných roli LE v ASR zanedbávala důležitost komunikace při vzniku LE a analyzovala data od mluvčích, kteří četli text v hluku, aniž by jim byla poskytnuta zpětná vazba o tom, zda je jejich projev srozumitelný. V této práci je pro sběr Lombardových promluv použit nový systém, zajišťující přítomnost komunikačního faktoru. Analýza získaných nahrávek ukázala významné rozdíly mezi neutrálními a LE daty pro řadu parametrů řečové produkce. Přechod z neutrální na LE řeč v ASR experimentech způsobil podstatné zhoršení úspěšnosti rozpoznávání v úlohách s velkým i malým slovníkem.

Druhá část disertace se zaměřuje na návrh metod omezujících dopad LE na ASR, založených na *ekvalizaci LE*, *robustních parametrizacích* a *modifikacích modelů*. Cílem *ekvalizace LE* je transformace příznaků Lombardovy řeči směrem k neutrální ještě před jejím zasláním akustickým modelům. Pro tento účel byly navrženy algoritmy modifikované normalizace vokálního traktu a formanty řízeného borcení frekvenční osy. Obě metody výrazně zvýšily přesnost rozpoznávání řeči pod LE. Další, částečně úspěšný, způsob ekvalizace LE byl realizován komerčním systémem hlasové konverze. Na základě daty řízeného návrhu byla získána sada *robustních parametrizací*. Parametrizace byly zkonstruovány nahrazením bank filtrů v MFCC (mel-frekvenční kepstrální koeficienty) a PLP (perceptuální lineární predikce) bankami lépe korespondujícími s rozložením lingvistické informace ve frekvenci. Použitím nových parametrizací v ASR systému bylo dosaženo podstatného zlepšení odolnosti vůči LE v porovnání se standardními MFCC a PLP parametrizacemi a state-of-the-art MR-RASTA (multi-resolution relative spectra) a Expolog parametrizacemi. V oblasti *modifikace modelů* se ukázala adaptace akustických modelů jako vhodný prostředek redukce rozdílů mezi neutrálními charakteristikami modelovanými v ASR systému a parametry Lombardovy řeči. Na závěr byl navržen dvoustupňový rozpoznávací systém (TSR) sestávající z klasifikátoru stylu řeči (neutrální/LE) a stylově-specifických akustických modelů. V porovnání s vícestupňovými systémy prezentovanými v literatuře TSR pro trénovaní stylově-specifických modelů postačuje pouze neutrální řeč. V ASR úloze na směsi neutrálních a Lombardových promluv TSR výrazně překonal diskrétní stylově-specifické rozpoznávače. Přínosy předkládané práce rozšiřují dosavadní poznatky o Lombardově efektu a přispívají k vývoji robustního rozpoznávání řeči v hlučném prostředí.

Acknowledgments

I would like to thank my thesis advisor, Petr Pollák, for his guidance and support throughout the course of my doctoral research, for offering me a great research topic, giving me the freedom to explore a variety of areas related to it, and for cooking the best home-made sausages.

I express my deep gratitude to Pavel Sovka for his enthusiastic lectures, which initiated and fostered my interest in the field of speech signal processing, and for his kindness and support as my boss and as the head of the Department of Circuit Theory. Among other faculty and staff at CTU in Prague, I would like to thank Roman Čmejla for unveiling to me the mysteries of speech production mechanisms, Rado Bortel for our discussions on math-related peculiarities, Václav Hanžl for taking care of the Linux cluster Magi, without which most of the figures and tables presented in this thesis would not exist, and the 'acoustic people' Libor Husník, Ondřej Jiříček, and Marek Brothánek, who were always willing to share their knowledge and lab equipment.

I am very grateful to Harald Höge from Siemens Corporate Technology for initiating the research project 'Normalization of Lombard Effect', for raising questions that never had simple answers, and for the funding that allowed me to finish this work. I would also like to thank David Sündermann from Siemens/SpeechCycle for his excellent collaboration within the Siemens project and for being a good friend. I would like to express my sincere gratitude to John H. L. Hansen, whose studies spanning over two decades, as well as our discussions at conferences were very inspiring for my work. I also thank him for offering me a position in the Center for Robust Speech Systems at the University of Texas at Dallas, and for giving me a generous portion of time to finish the thesis writing.

I would like to thank Jan Nouza from the Technical University of Liberec for providing me with TUL's state-of-the-art ASR framework for large vocabulary experiments. In particular, this work has also benefited from collaborations with my colleagues and friends Petr Fousek, Jindřich Žďánský, and Petr Červa. I would like to thank my lab-mate Petr Fousek for our lunch DSP discussions, for making everything parallel, for pretending that Linux has also a human side, and for recording amazing bass lines. I am thankful to my Liberec colleagues Jindřich Žďánský and Petr Červa for sharing their frameworks, and for teaching me that coffee and a cool head are solutions to almost every problem.

Many thanks to my friends Zoraida Callejas Carrión, CRSS, Darja Fišer, Milan Kníže, Hynek Kocourek, Václav Mocek, Jan Novotný, Petr Prášek, Radek Skarnitzl, Petra–Maria Strauß, Pavel Štemberk, Jiří Tatarinov, and Jan Volín for their support, patience, kindness, and for making life both inside and outside of the lab special. I am also grateful to Keith Godin and Hynek Kocourek for being my English language tutors and for proofreading large portions of the manuscript.

Finally, I would like to thank my family for their unconditional support, my brother Tomáš for coding the H&T Recorder, my parents for providing me with home where I can always return, and my grandparents for continuously spoiling me.

This research was supported in part by the following grants and research activities: COST 278 'Spoken Language Interaction in Telecommunication', GAČR 102/05/0278 'New Trends in Research and Application of Voice Technology', MSM 6840770014 'Research in the Area of the Prospective Information and Navigation Technology', 1ET201210402 'Voice Technologies in Information Systems', and GAČR 102/03/H085 'Biological and Speech Signals Modeling'.

Contents

List of Figures

List of Tables

List of Abbreviations

AC	Autocorrelation
AMDF	Average Magnitude Difference Function
ANN	Artificial Neural Network
ASR	Automatic Speech Recognition
BL	Baseline
CC	Cross-Correlation
CELP	Code-Excited Linear Prediction
CFV	Classification Feature Vector
CLE	Converted LE Speech, both excitation and formants transformed
CLEF_0	Converted LE Speech, only excitation transformed
CLSD‘05	Czech Lombard Speech Database ‘05
CS	Condition-Specific
CTU	Czech Technical University in Prague
CV	Cross-Validation
CZKCC	Temic Czech Car Database
DCT	Discrete Cosine Transform
DFE	Discriminative Feature Extraction
DFT	Discrete Fourier Transform
DM	Diagonal Covariance Matrix
DTFE	Direct Time Domain Fundamental Frequency Estimation
DTW	Dynamic Time Warping
ECESS	European Center of Excellence on Speech Synthesis
EIH	Ensemble Interval Histogram
ELRA	European Language Resource Association
EM	Expectation-Maximization
FB	Filter Bank
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
FM	Full Covariance Matrix
FWS	Full-Wave Spectral Subtraction
GD	Gender-Dependent or Group-Dependent
GMLC	Gaussian Maximum Likelihood Classifier
GMM	Gaussian Mixture Model
HFCC	Human Factor Cepstral Coefficients
HMM	Hidden Markov Model
HTK	Hidden Markov Model Toolkit
ID	Information Distribution
IDCT	Inverse Discrete Cosine Transform
IIR	Infinite Impulse Response

ITU	International Telecommunication Union
LCD	Liquid Crystal Display
LDC	Linguistic Data Consortium
LE	Lombard Effect
LFCC	Linear Frequency Cepstral Coefficients
LM	Language Model
LP	Linear Prediction
LPC	Linear Predictive Coding
LSF	Line Spectrum Frequencies
LVCSR	Large Vocabulary Continuous Speech Recognition
MAP	Maximum A Posteriori Estimation
MFCC	Mel Frequency Cepstral Coefficients
ML	Maximum Likelihood
MLLR	Maximum Likelihood Linear Regression
MLP	Multi-Layer Perceptron
MR–RASTA	Multi-Resolution RASTA
MVDR	Minimum Variance Distortionless Response
NS	Noise Suppression
OL	Open-Loop
OLA	Overlap-and-Add
PCM	Pulse Code Modulation
PDA	Pitch Detection Algorithm
PDF	Probability Density Function
PLP	Perceptual Linear Prediction
PMA	Pitch Marking Algorithm
PMC	Parallel Model Combination
RAPT	Robust Algorithm for Pitch Tracking
RASTA	Relative Spectral Transformation
RFCC	Repartitioned Filter Bank Cepstral Coefficients
RMS	Root Mean Square
SD	Speaker-Dependent
SHS	Sub-Harmonic Summation
SI	Speaker-Independent
SNR	Signal-to-Noise Ratio
SoX	Sound Exchange Tool
SPEECON	Speech Driven Interfaces for Consumer Applications
SPINET	Spatial Pitch Network
SPL	Sound Pressure Level
SUSAS	Speech under Simulated and Actual Stress
TEO	Teager Energy Operator
TSR	Two-Stage Recognition System
TUL	Technical University of Liberec
UE	Unvoiced Error
UER	Utterance Error Rate
UT-Scope	Database of Speech under Cognitive and Physical Stress and Emotion
V/UV	Voiced/Unvoiced
VAD	Voice Activity Detector
VE	Voiced Error
VTL	Vocal Tract Length
VTLN	Vocal Tract Length Normalization
WER	Word Error Rate

Chapter 1

Introduction

Speech is one of the most advanced means of conveying thoughts, desires, and emotions between human beings. Currently, due to the remarkable progress in the field of speech technology and natural language processing, speech has become also a convenient and effective component of human–machine interaction, spanning application domains such as voice control and dictation, information retrieval, and travel arrangements. Moreover, automatic speech processing plays an important role in the transcription of broadcast news and sound archives, and in speech-to-speech translation, which performs a real-time interpretation from one spoken language to another. The last case demonstrates how advances in speech and language technology not only improve communication between humans and machines, but also between humans.

One of the key stages in speech-oriented applications is extraction of linguistic information from the acoustic speech signal and its conversion into text representation. This task, called automatic speech recognition (ASR), is handled by speech recognizers. An ideal speech recognizer transforms the speech signal into the corresponding text regardless of who is talking and what is being said, independent of the environment where the speech is produced, as well as the type of channel used to deliver the speech signal to the recognizer. In spite of the relative success of the last decades in transferring ASR technology from laboratory conditions to the real-world tasks, state-of-the-art recognizers are still very sensitive to disfluencies in speech and to changes in environmental and speaker characteristics from those considered during the system design. In particular, the following factors strongly impact recognition performance:

- *Environmental variability*: additive noises, convolutional distortions (reverberation, microphone/channel characteristics), (Junqua, 2002),

- *Speaker variability*: gender (Olsen and Dharanipragada, 2003), age (Blomberg and Elenius, 2004), dialect (Clarke and Jurafsky, 2006), and stress (environmental, emotional, workload), (Yapanel and Hansen, 2008),

- *Disfluencies in speech*: editing terms ('*oops*'), repetitions, revisions (content replacement), restarts, fragments, filled pauses, and discourse markers ('*Well, you know, I mean, ...*'), (Liu *et al.*, 2006).

Numerous research groups within the speech processing community are searching for ASR solutions more resistant to these factors. This thesis attempts to contribute to these efforts, focusing in particular on the impact of noise-induced environmental stress known as Lombard effect.

1.1 Objective

In a noisy environment, speakers tend to adjust the way they talk in order to maintain intelligible communication over noise, (Junqua, 1993), (Hansen, 1996). This phenomenon is called Lombard effect (LE) after the French oto-rhino-laryngologist Etienne Lombard, who first described the impact of noise on speech production, (Lombard, 1911). The speech production variations due to LE may considerably deteriorate the performance of ASR systems trained on neutral speech (modal speech produced in a calm environment), (Bou-Ghazale and Hansen, 2000). The ASR accuracy degradation by LE can be significantly stronger than the one caused by the presence of background noise in the speech signal, (Rajasekaran *et al.*, 1986), (Takizawa and Hamada, 1990).

The goal of this thesis is to analyze differences in the production of neutral speech and speech under LE (Lombard speech), and to propose algorithms increasing resistance of ASR systems to LE. All feature analyses and ASR experiments are conducted on Czech speech corpora. It is noted that previous studies have not considered the Czech language and therefore this represents a new advancement.

1.2 Motivation

1.2.1 Analysis of speech under LE

During the last hundred years, a number of studies have analyzed the impact of Lombard effect on speech production. Considerable production differences between neutral speech and speech produced under Lombard effect were observed across studies, see Sec. 3.2 for a detailed overview. Unfortunately, at the level of particular speech parameters, the reported trends often disagree, (Womack and Hansen, 1999). This is presumably due to the fact that a majority of the analyses employed only a very limited number of utterances from a few subjects. The number of participating speakers was ranging typically from ten (Webster and Klumpp, 1962), (Lane *et al.*, 1970), (Junqua, 1993), to one or two speakers, (Summers *et al.*, 1988), (Pisoni *et al.*, 1985), (Bond *et al.*, 1989), (Tian *et al.*, 2003), (Garnier *et al.*, 2006). Another reason for the variability of the observations lies in differences between the experimental setups used. With a few exceptions, a communication factor has not been involved in the data collections. Speakers only read utterances without a need to convey the message over noise to a listener, (Junqua *et al.*, 1999). Here, the motivation for the speaker's reactions to noise was not clearly defined, hence, the resulting production changes were strongly time-varying and speaker-dependent.

Although a majority of analyses focused on English speech corpora, several studies considered also other languages, e.g., Spanish (Castellanos *et al.*, 1996), French (Garnier *et al.*, 2006), Japanese (Wakao *et al.*, 1996), Korean, (Chi and Oh, 1996), or Mandarin Chinese (Tian *et al.*, 2003). So far, no systematic research on LE in Czech speech has been conducted. The aim of this thesis is to analyze speech production differences in Czech neutral and Lombard speech and use this knowledge for the design of LE-robust ASR algorithms. To address some of the LE analysis issues occurring in past studies, the following steps are conducted:

- Proposal of a data-collection setup imposing a communication factor into the recording,
- Acquisition of a Czech speech database comprising an extensive set of neutral and Lombard speech utterances from a larger number of subjects.

1.2.2 Automatic Recognition of Lombard Speech

Acoustic models in state-of-the-art ASR systems are typically trained on neutral speech data. Speech production changes introduced by LE cause a mismatch between neutral-trained models and

the processed speech, often resulting in a severe deterioration of ASR performance. Despite the fact that the impact of LE on ASR is known to the speech community, it receives only limited attention in the design of state-of-the-art recognition systems.

The efforts to increase the ASR resistance to LE can be divided into the following groups (a detailed overview is provided in Sec. 3.5):

- Robust features – a search for speech signal representations less sensitive to LE, (Hanson and Applebaum, 1990a), (Bou-Ghazale and Hansen, 2000), (Yapanel and Hansen, 2008),

- LE-equalization – transformations of Lombard speech parameters towards neutral, (Hansen and Bria, 1990), (Takizawa and Hamada, 1990), (Suzuki *et al.*, 1994), (Hansen, 1996), (Bou-Ghazale and Hansen, 2000),

- Model adjustments – transformations of neutral-trained acoustic models towards condition-specific models, (Gauvain and Lee, 1994), (Gales and Woodland, 1996), novel acoustic-modeling architectures, (Womack and Hansen, 1999),

- Training methods – merging samples from a variety of conditions into a single training set to obtain condition-independent models (multistyle training), (Lippmann *et al.*, 1987), (Yao *et al.*, 2004), training condition-specific acoustic models, (Bou-Ghazale and Hansen, 1998).

The approaches based on improved *training methods* assume that there is a sufficient amount of data available for modeling the condition-specific characteristics. However, due to a strong variability of LE with the level and type of background noise, only a limited number of Lombard speech samples may be available for adapting the real-world system to the actual conditions. Here, *model adjustment* techniques employing acoustic model adaptation are preferable.

Although some of the methods proposed in the past provide substantial improvement to Lombard speech recognition, the attained error rates are still considerably higher than those for neutral speech. The goal of the present thesis is to design algorithms that further improve performance of ASR under Lombard effect, with a major focus on the domains of *robust features*, *LE-equalization*, and *model adjustments*.

1.3 Original Contributions

The primary focus of this work is in the analysis and equalization of Lombard effect in the Czech spoken language. While previous studies on LE considered several widely spoken languages, no research has been conducted on the Czech speech. An analysis of commercial Czech speech corpora conducted in the initial stage of this thesis has shown that in the available speech data, in spite of being acquired in adverse noisy conditions, the occurrence of LE is very sparse (see Chap. 6). Hence, a new speech corpus, Czech Lombard Speech Database (CLSD'05), was designed and acquired. To assure the presence of LE in CLSD'05, a setup motivating speakers to maintain intelligible communication over simulated background noise was designed and used in the acquisition of the Lombard recordings. It is noted that a majority of previous studies on the impact of LE on ASR have ignored the importance of the communication loop in evoking Lombard effect, and instead analyzed data from subjects who read text in noise without being provided feedback regarding whether their speech was intelligible (see Sec. 3.6). To further advance the simulated LE recording setup for the purpose of future recording, auditory feedback attenuation due to the use of closed headphones was measured and its compensation was proposed. The major contributions of this thesis are summarized below:

- Novel pitch detection algorithm: Computationally efficient time-domain pitch tracker is proposed. The algorithm is evaluated side by side with five state-of-the-art pitch trackers on the ECESS reference database.

- Acoustic model adaptation: Efficiency of speaker dependent/independent adaptation of acoustic models to Lombard effect is tested[1]. A considerable recognition performance improvement is reached when applying both speaker-dependent and speaker-independent model adaptation.

- Voice conversion: Voice conversion framework is trained on parallel Lombard/neutral utterances and used for normalizing Lombard speech towards neutral[2]. Voice conversion is included in the ASR front-end, yielding an improvement in small vocabulary recognition task on LE speech.

- Data-driven design of robust features: Contribution of frequency sub-bands to speech recognition performance is studied. New front-end filter banks for MFCC and PLP-based front-ends are designed, providing superior robustness to LE. It is shown that filter banks inspired by auditory models do not represent the optimal choice for ASR front-ends.

- Vocal tract length normalization: Modified vocal tract length normalization scheme is proposed. Impact of limiting the speech bandwidth on distribution and evolution of frequency warping factors is studied. Efficiency of speaker and utterance driven warping is compared, both of them providing substantial improvement of recognition performance.

- Formant-driven frequency warping: A function mapping average LE formant locations to average neutral ones is determined and incorporated in the ASR front-end, considerably increasing accuracy of the LE speech recognition. Surprisingly, the LE–neutral formant transformation does not significantly deteriorate recognition performance when applied also to neutral speech.

- Classification of neutral/LE speech: Based on the speech feature distributions found in neutral and LE speech, set of gender/lexicon-independent parameters efficient for neutral/LE classification is proposed. Discriminative properties of spectral slopes extracted from various frequency bands of short-time spectra are studied in detail.

- Two-stage recognition system (TSR): An ASR system for neutral/LE speech recognition is designed. In the first stage, classifier decides whether the incoming utterance is neutral or Lombard. In the second stage, the utterance is passed to the corresponding neutral-specific or LE-specific recognizer. When exposed to the mixture of neutral and LE utterances, TSR significantly outperforms both neutral-specific and LE-specific recognizers. Acoustic models of TSR require only neutral speech samples for training.

1.4 Thesis Outline

This chapter briefly discussed the automatic speech recognition (ASR) issues introduced by the presence of environmental noise, with a particular focus on the impact of Lombard effect. Subsequently, goals and outcomes of the thesis were presented. The following chapters of the thesis are organized as follows.

[1]The framework was provided and the model adaptation conducted by Dr. Petr Červa, Technical University of Liberec. Author of the thesis designed the experiments and provided data for adaptation.

[2]Using voice conversion (VC) for normalization of Lombard speech was proposed by Prof. Harald Höge, Siemens Corporate Technology, Munich, Germany. David Sündermann (Siemens) provided the VC system (Sündermann *et al.*, 2006b) and conducted the system training and data conversion. Author of the thesis provided data for VC training and for conversion, analyzed impact of VC on speech features, and evaluated VC efficiency in the ASR experiments.

Chapter 2 defines the task of ASR and describes the algorithms and typical structures used in the HMM-based ASR systems. The principles and terms introduced in this chapter form the background for the majority of algorithm formulations and experiments conducted throughout the thesis.

Chapter 3 discusses the impact of Lombard effect on the speech production parameters, and presents an overview of the state of the art techniques for stress and talking style classification, robust speech recognition, and Lombard corpora acquisition.

Chapter 4 describes the framework for feature analyses and speech recognition, as well as evaluation metrics used in the thesis experiments. In addition, a novel algorithm for pitch extraction is proposed and compared to the state of the art pitch trackers on a reference database.

Chapter 5 details the design, recording setup, and content of the Lombard speech database acquired for the purposes of the thesis algorithm development and experiments. Auditory feedback attenuation caused by wearing headphones during database recording is analyzed and a speech feedback compensation for the attenuation is proposed.

Chapter 6 explores the suitability of selected Czech speech corpora for the study of Lombard effect. For each of the speech databases, feature analyses and recognition tasks are conducted on neutral speech and speech uttered in noise.

In *Chapter 7*, acoustic model adaptation is used to transform neutral speaker-independent models towards Lombard speech characteristics. The effectiveness of speaker-dependent and speaker-independent adaptation is compared.

In *Chapter 8*, voice conversion is applied to normalize Lombard speech towards neutral. Parameters of the normalized speech are compared to the actual neutral speech samples to evaluate the accuracy of the conversion. Subsequently, voice conversion is incorporated into the ASR front-end and evaluated in the recognition tasks.

Chapter 9 presents a novel approach to the design of feature extraction filter banks. In the data-driven procedure, bands of the initial filter bank are redistributed according to the observed linguistic information distribution. The procedure is used to design feature extraction front-ends with improved resistance to Lombard effect. The performance of the newly proposed front-ends is compared to the traditional methods in the set of ASR tasks employing changes of talking style (neutral, LE), average fundamental frequency, and noisy background.

In *Chapter 10*, modified vocal tract normalization and formant-based frequency warping are proposed. The goal here is to address formant shifts due to Lombard effect by warping the speech spectra. Both algorithms are included into the ASR front-end and evaluated in the recognition tasks.

Chapter 11 establishes a set of speech features effective for neutral/LE classification. Subsequently, a novel two-stage recognition system comprising neutral/LE classifier and style specific recognizers is formulated and evaluated on a collection of neutral and Lombard utterances.

Chapter 12 summarizes major findings and contributions of the thesis and discusses possible future research directions in the field of LE-robust ASR.

Chapter 2

ASR Background

Early ASR systems were based on knowledge-driven algorithms. System designers were establishing rules of mapping the acoustic signal to phonetic units, as well as higher level knowledge (lexicon, syntax, semantics, and pragmatics). The recognition process comprises two steps. First, segmentation and labeling of the acoustic signal is carried out. Based on the time evolution of the acoustic features, the signal is divided into discrete regions, each being assigned one or several candidate phonetic labels. In the second step, artificial intelligence is applied at the higher level knowledge in the extraction of word strings, (Klatt, 1977). These systems are computationally demanding and reach poor recognition results, particularly due to the difficulty in establishing a set of rules for reliable acoustic-phonetic matching, Rabiner and Juang (1993).

In the last thirty years, pattern-matching techniques became preferred in ASR. Here, the relations between spoken and transcribed speech are searched in the process of training the ASR system on speech corpora. If there is a sufficient amount of realizations of each of the classes to be recognized (e.g., phones, words, phrases), the training procedure captures characteristic properties (patterns) of each class. In the recognition stage, the speech is compared to the stored class patterns and classified based on the goodness of match. Two main approaches to pattern matching have been widely used in ASR – deterministic pattern matching based on dynamic time warping (DTW) (Sakoe and Chiba, 1978), and stochastic pattern matching employing hidden Markov models (HMMs) (Baker, 1975).

In DTW, each class to be recognized is represented by one or several templates. Using more than one reference template per class may be preferable in order to improve the pronunciation/speaker variability modeling. During recognition, a distance between an observed speech sequence and class patterns is calculated. To eliminate the impact of the duration mismatch between test and reference patterns, stretched and warped versions of the reference patterns are also employed in the distance calculation. The recognized word corresponds to the path through the model that minimizes the accumulated distance. Increasing the number of class pattern variants and loosening warping constrains may improve DTW-based recognition performance at the expense of storage space and computational demands.

In state of the art systems, HMM-based pattern matching is preferred to DTW due to better generalization properties and lower memory requirements. The speech recognition frameworks considered in this thesis employed exclusively HMMs. A typical structure of an HMM-based ASR system is shown in Fig. 2.1. The remainder of this chapter will discuss the individual stages of HMM-based recognition in detail.

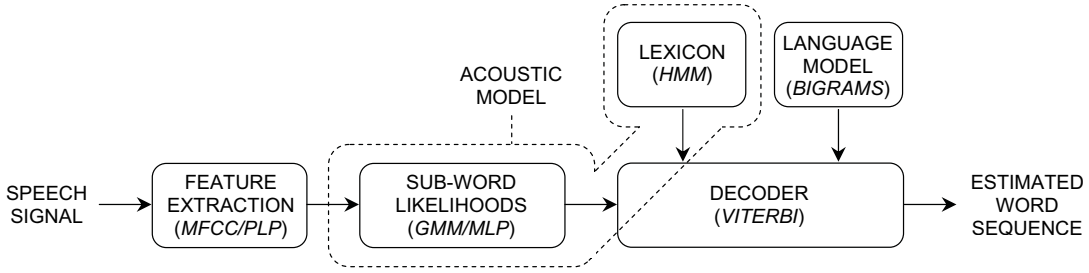

Figure 2.1: Block diagram of typical HMM-based automatic speech recognition system.

2.1 ASR Definition

The goal of automatic continuous speech recognition is to find the most likely sequence of words $\hat{\boldsymbol{W}} = w_1, w_2, \ldots, w_N$ in the language $\mathcal{L}$ given an acoustic input $\boldsymbol{O}$ comprising a sequence of parameterized acoustic observations $\boldsymbol{O} = \boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T$. This problem can be expressed probabilistically as follows:

$$\hat{\boldsymbol{W}} = \arg\max_{\boldsymbol{W} \in \mathcal{L}} P\left(\boldsymbol{W} | \boldsymbol{O}, \boldsymbol{\Theta}\right), \tag{2.1}$$

where $\boldsymbol{\Theta}$ is a set of model parameters. It is problematic to estimate the probability $P\left(\boldsymbol{W} | \boldsymbol{O}, \boldsymbol{\Theta}\right)$ directly, hence Bayes' rule is applied:

$$\hat{\boldsymbol{W}} = \arg\max_{\boldsymbol{W} \in \mathcal{L}} \frac{P\left(\boldsymbol{O} | \boldsymbol{W}, \boldsymbol{\Theta}\right) P\left(\boldsymbol{W} | \boldsymbol{\Theta}\right)}{P\left(\boldsymbol{O} | \boldsymbol{\Theta}\right)}. \tag{2.2}$$

Because the term $P\left(\boldsymbol{O} | \boldsymbol{\Theta}\right)$ in Eq. (2.2) is constant for all hypotheses $\boldsymbol{W} \in \mathcal{L}$, it can be omitted. The set of model parameters $\boldsymbol{\Theta}$ may comprise two components, $\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_a, \boldsymbol{\Theta}_l\}$, where $\boldsymbol{\Theta}_a$ are parameters of the acoustic model, and $\boldsymbol{\Theta}_l$ are parameters of the language model. Assuming that $\boldsymbol{\Theta}_a$ and $\boldsymbol{\Theta}_l$ are independent, and that the prior probability of the word sequence $P\left(\boldsymbol{W} | \boldsymbol{\Theta}\right)$ will be determined exclusively by the language model, while the probability of the sequence of acoustic observations given the sequence of words $P\left(\boldsymbol{O} | \boldsymbol{W}, \boldsymbol{\Theta}\right)$ will be determined exclusively by the acoustic model, (2.2) can be rewritten as:

$$\hat{\boldsymbol{W}} = \arg\max_{\boldsymbol{W} \in \mathcal{L}} P\left(\boldsymbol{O} | \boldsymbol{W}, \boldsymbol{\Theta}_a\right) P\left(\boldsymbol{W} | \boldsymbol{\Theta}_l\right). \tag{2.3}$$

Using this equation, the acoustic model and language model survey can be derived separately in the following two sections.

2.2 Acoustic Model

To find parameters of the acoustic model, a training set comprising acoustic sequences and their transcriptions is required. The parameters are estimated so as to maximize the probability of the training observation vectors:

$$\hat{\boldsymbol{\Theta}}_a = \arg\max_{\boldsymbol{\Theta}_a} \prod_{\{\boldsymbol{O}; \boldsymbol{W}\} \in \mathcal{T}} P\left(\boldsymbol{O} | \boldsymbol{W}, \boldsymbol{\Theta}_a\right), \tag{2.4}$$

where $\mathcal{T}$ represents the set of training utterances and their transcriptions. Since it is not possible to find the global maximum of the probability for $\hat{\boldsymbol{\Theta}}_a$ analytically, iterative procedures are used to find the local maxima.

Recently, hidden Markov models (HMMs) have been commonly used for acoustic modeling. In contrast with Markov models, where to each state of the model an observable event is assigned, in HMMs, the observation is a probabilistic function of the state. Here, only the resulting observations generated by the emission probabilities of the states can be seen, while the level of contribution of the individual states, and thus the sequence of states that would most likely generate the observations, remains hidden. The HMM model is specified by:

- Set of states $\boldsymbol{Q} = \{q_i\}$.
- Set of emission probability density functions $\boldsymbol{B} = \{b_j(\boldsymbol{o}_t)\}$, where $b_j(\boldsymbol{o}_t)$ is a probability density of $\boldsymbol{o}_t$ being generated by state j.
- Transition probability matrix $\boldsymbol{A} = \{a_{ij}\}$, where a_{ij} is the probability of the transition from state i to state j:
$$a_{ij} = P(q_{t+1} = j \,|\, q_t = i), \tag{2.5}$$
where q_t represents the actual state in time t, and i, j states of the HMM.
- Initial state occupancies $\boldsymbol{\pi} = \{\pi_j\}$, which represent the probabilities that state j is the initial state, $\pi_j = P(q_1 = j)$.

There are two popular approaches to modeling observation likelihoods – Gaussian probability density functions (PDFs), and multi-layer perceptrons (MLPs), Rabiner and Juang (1993), Jurafsky and Martin (2000).

- Gaussian probability density function (PDF): the observation probability $b_j(\boldsymbol{o}_t)$ for an n-dimensional vector $\boldsymbol{o}_t$ is represented by a multi-variate Gaussian function. To better capture the distribution of the observation samples, multivariate multiple-mixture Gaussian distributions are often used:
$$b_j(\boldsymbol{o}_t) = \sum_{m=1}^{M} c_{jm} \mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}\right), \tag{2.6}$$
where M is the number of mixtures, c_{jm} is the weight of the m-th mixture component of the j-th state so that
$$\sum_{m=1}^{M} c_{jm} = 1, \tag{2.7}$$
and $\mathcal{N}$ is a multi-variate Gaussian with mean vector $\boldsymbol{\mu}_{jm}$ and covariance matrix $\boldsymbol{\Sigma}_{jm}$:
$$b_{jm}(\boldsymbol{o}_t) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}_{jm}|}} \cdot \mathrm{e}^{-\frac{1}{2}(\boldsymbol{o}_t - \boldsymbol{\mu}_{jm})^T \boldsymbol{\Sigma}_{jm}^{-1} (\boldsymbol{o}_t - \boldsymbol{\mu}_{jm})}. \tag{2.8}$$
Here, the state emission PDFs are called Gaussian mixture models (GMMs). It is often assumed that the components of $\boldsymbol{o}_t$ are independent. In this case, the full covariance matrix $\boldsymbol{\Sigma}_{jm}$ is replaced by a diagonal matrix comprising only variances of the components of $\boldsymbol{o}_t$.
- Multi-layer perceptron[1] – an artificial neural network comprising a set of computation units (neurons) connected by weighted links (synapses). In HMMs, MLP has one output for each state j. The sum of outputs is kept equal to '1'. The network estimates a probability of

[1]The typical structure of the multi-layer perceptron and the process of its training are further discussed in Sec. 11.1.2.

an HMM state j given the observation vector $\boldsymbol{o}_t$, $P(j|\,\boldsymbol{o}_t)$. To get the observation likelihood $P(\boldsymbol{o}_t|\,j)$, which is needed for the HMM, Bayes' rule can be used:

$$P(j|\,\boldsymbol{o}_t) = \frac{P(\boldsymbol{o}_t|\,j)\,P(j)}{P(\boldsymbol{o}_t)}, \tag{2.9}$$

which can be rearranged as:

$$\frac{P(\boldsymbol{o}_t|\,j)}{P(\boldsymbol{o}_t)} = \frac{P(j|\,\boldsymbol{o}_t)}{P(j)}. \tag{2.10}$$

The numerator on the right side of (2.10) is the MLP's output assigned to the state j and the denominator is the prior probability of the state j, which can be determined from the training set by calculating the ratio of occurrences of j to the total sum of occurrences of all states. The ratio in Eq. (2.10) is called scaled likelihood. Since the probability $P(\boldsymbol{o}_t)$ is constant for all state probabilities, the scaled likelihood is as efficient as the regular likelihood $P(\boldsymbol{o}_t|\,j)$.

Given the HMM model, the probability of the state sequence $\boldsymbol{q} = q_1, q_2, \ldots, q_T$ being generated by the model is:

$$P(\boldsymbol{q}|\,\boldsymbol{A}, \boldsymbol{\pi}) = \pi_{q_0} a_{q_0 q_1} a_{q_1 q_2} \ldots a_{q_{T-1} q_T}. \tag{2.11}$$

The probability of a parameterized acoustic sequence $\boldsymbol{O}$ being generated by the HMM is defined:

$$P(\boldsymbol{O}|\,\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{Q}, \boldsymbol{\pi}) = \sum_{\boldsymbol{q}} \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1} q_t} b_{q_t}(\boldsymbol{o}_t). \tag{2.12}$$

If recognition tasks employ larger vocabularies, it would be very costly to build a unique acoustic model for every word in the lexicon. Hence, the words are modeled using sub-word units (e.g., phones). The size of such units can be considerably smaller than the size of the lexicon, since the units are common to all words. In this case, Equation (2.3) can be re-written as:

$$\hat{\boldsymbol{W}} = \arg\max_{\boldsymbol{W}} \left\{ \max_{\boldsymbol{W} \in L,\, \boldsymbol{U} \in \mathcal{U}} \left[P(\boldsymbol{O}|\,\boldsymbol{U}, \boldsymbol{\Theta}_a)\, P(\boldsymbol{U}|\,\boldsymbol{W}, \boldsymbol{\Theta}_p)\, P(\boldsymbol{W}|\,\boldsymbol{\Theta}_l) \right] \right\}, \tag{2.13}$$

where $\boldsymbol{U} = u_1, u_2, \ldots, u_M$ is the sequence of sub-word units, $\mathcal{U}$ is the set of all possible sub-word units' sequences defined by the pronunciation lexicon (dictionary of word pronunciations), $P(\boldsymbol{O}|\,\boldsymbol{U}, \boldsymbol{\Theta}_a)$ is the probability of the sequence of observations $\boldsymbol{O}$ given the sequence of sub-word units and the acoustic model, and $P(\boldsymbol{U}|\,\boldsymbol{W}, \boldsymbol{\Theta}_p)$ is the probability of the sequence of units given the sequence of words and so called pronunciation model $\boldsymbol{\Theta}_p$, representing the probability that $\boldsymbol{U}$ is a transcription of word w.

In the case of MLP-based sub-word models, each of the MLP outputs typically represents a scaled likelihood of a single context-independent phone (Hermansky and Fousek, 2005). Application of this approach will be demonstrated in Sec. 9.2.2. GMM-based models usually comprise multiple states to capture the time evolution of the sub-word units (phones, triphones), Young *et al.* (2000). Since no analytical solution to finding model parameters as mentioned in Eq. (2.4) is known, iterative procedures such as the Baum-Welch forward-backward expectation-maximization (EM) algorithm in case of GMMs or error back-propagation in MLPs are used, Rabiner and Juang (1993).

Based on the pronunciation lexicon, the sub-word unit models are connected to form the word pronunciation models. These models form a lattice of the joint 'acoustic-phonetic' HMM. Probabilities of transitions between word models are driven by the language model, which is discussed in the following section.

2.3 Language Model

The language model $P(\boldsymbol{W})$, given its parameters $\boldsymbol{\Theta}_l$, estimates the probability of a word sequence $w_1, w_2, \ldots, w_N$. This probability can be calculated using the chain rule, Papoulis (2001):

$$P(\boldsymbol{W}) = P(w_1) \prod_{i=2}^{N} P(w_i | w_1, w_2, \ldots, w_{i-1}). \tag{2.14}$$

However, it is not possible to find and store probabilities for all possible sequences of words, hence (2.14) cannot be directly calculated. One of the ways to estimate probability of the word sequence is using an n-gram model. In the n-gram model, the history of all previously uttered words is approximated by the n most recent words:

$$P(w_i | w_1, w_2, \ldots, w_{i-1}) \approx P(w_i | w_{i-n+1}, w_{i-n+2}, \ldots, w_{i-1}). \tag{2.15}$$

Thus:

$$P(w_1, w_2, \ldots, w_N) \approx P(w_1, w_2, \ldots, w_{n-1}) \prod_{i=n}^{N} P(w_i | w_{i-n+1}, w_{i-n+2}, \ldots, w_{i-1}). \tag{2.16}$$

As shown in the first term on the right side of (2.16), to estimate the probability of the word sequence using n-grams, $n-1, n-2, \ldots,$ 2-grams and prior probabilities $P(w_i)$ of each word must be known. In the LVCSR experiments presented in the following chapters, a bigram language model was used. In this case, the estimation of probability of word sequence reduces to the calculation:

$$P(w_1, w_2, \ldots, w_N) \approx P(w_1) \prod_{i=2}^{N} P(w_i | w_{i-1}). \tag{2.17}$$

The right side of (2.17) represents parameters of the bigram model (i.e., word priors and bigram probabilities). The prior probabilities can be estimated:

$$P(w_i) = \frac{count(w_i)}{\sum_{w \in Lex} count(w)}, \tag{2.18}$$

where $\sum_{w \in Lex} count(w)$ represents the total sum of occurrences of word w from the lexicon Lex in the training corpus. The bigram probabilities can be estimated:

$$P(w_i | w_{i-1}) = \frac{count(w_{i-1}, w_i)}{\sum_{w_j, w_k \in Lex} count(w_j, w_k)}, \tag{2.19}$$

where $count(w_{i-1}, w_i)$ is a number of representations of the bigram w_{i-1}, w_i and the denominator in Eq. (2.19) is number of representations of all bigrams in the training corpus.

2.4 Decoding

In the decoding stage, as shown in Eq. (2.13), the task is to find the most likely word sequence $\boldsymbol{W}$ given the observation sequence $\boldsymbol{O}$, and the 'acoustic-phonetic-language' model. The decoding problem can be solved using dynamic programming algorithms. Rather than evaluating likelihoods of all possible model paths generating $\boldsymbol{O}$, the focus is on finding a single path through the network

yielding the best match to $\boldsymbol{O}$. To estimate the best state sequence $\boldsymbol{q} = q_1, q_2, \ldots, q_T$ for the given observation sequence, the Viterbi algorithm is frequently used, Rabiner and Juang (1993), Young *et al.* (2000). Let

$$\delta_t(i) = \max_{q_1, q_2, \ldots, q_{t-1}} P(\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_t | q_1, q_2, \ldots, q_{t-1}, q_t = i, \boldsymbol{\Theta}) \tag{2.20}$$

be the highest probability along a single path ending at time t in state i. Suppose $\delta_t(i)$ is known, $\delta_{t+1}(j)$ can be calculated as:

$$\delta_{t+1}(j) = \max_i [\delta_t(i) a_{ij}] \cdot b_j(\boldsymbol{o}_{t+1}). \tag{2.21}$$

The path searching is then conducted as follows. First, given the observation $\boldsymbol{o}_1$, the initial probabilities of starting in state i are calculated for all states:

$$\delta_1(i) = \pi_i b_i(\boldsymbol{o}_1), \ 1 \leqslant i \leqslant N, \tag{2.22}$$

where N denotes a number of model states. Let $\boldsymbol{\psi}$ be a matrix storing state indices of the N path candidates:

$$\boldsymbol{\psi} = (\psi_{i,t})_{N \times T}, \tag{2.23}$$

where T is the length of the observation vector. The matrix is initialized:

$$\boldsymbol{\psi}(i, 1) = 0. \tag{2.24}$$

In the second stage, for each model state, the most probable evolution of the path is recursively estimated given the previous $\delta_{t-1}(i)$ for all states:

$$\delta_t(j) = \max_{1 \leqslant i \leqslant N} [\delta_{t-1}(i) a_{ij}] \cdot b_j(\boldsymbol{o}_t), \ 2 \leqslant t \leqslant T, \ 1 \leqslant j \leqslant N. \tag{2.25}$$

The matrix $\boldsymbol{\psi}$ is updated:

$$\boldsymbol{\psi}(j, t) = \arg\max_{1 \leqslant i \leqslant N} [\delta_{t-1}(i) a_{ij}], \ 2 \leqslant t \leqslant T, \ 1 \leqslant j \leqslant N. \tag{2.26}$$

When $\boldsymbol{o}_T$ is reached, the forward pass is terminated and the state in which the most probable path ends is picked as the most probable word sequence:

$$P = \max_{1 \leqslant i \leqslant N} [\delta_T(i)], \tag{2.27}$$

$$q_T = \arg\max_{1 \leqslant i \leqslant N} [\delta_T(i)]. \tag{2.28}$$

Subsequently, using $\boldsymbol{\psi}$, a path backtracking is conducted, in which the optimal path is traced from the final state q_T to the beginning and the overall probability of the path is calculated. To reduce the problems in finite precision arithmetic (i.e., all probabilities are less than one and multiplications cause very small numbers), multiplications in the Viterbi algorithm are usually replaced by summations of logarithms.

In the case of larger vocabulary recognition tasks, it would be challenging to consider all possible words during the recursive part of the Viterbi algorithm. To address this, a beam search can be used. Here, in each Viterbi iteration, only the words with path probabilities above a threshold are considered when extending the paths to the next time step. This approach speeds up the searching process at the expense of decoding accuracy.

The Viterbi algorithm assumes that each of the best paths at time t must be an extension of each of the best paths ending at time $t-1$, which is not generally true. The path that seems to be less probable than others in the beginning may turn into being the best path for the sequence as a whole (e.g., the most probable phoneme sequence does not need to correspond to the most probable word sequence). This issue is addressed by extended Viterbi and forward-backward algorithms, Jurafsky and Martin (2000).

2.5 Feature Extraction

The goal of feature extraction in ASR is to transform the speech signal into a parametric representation of reduced dimensionality, providing a good discriminability between classes to be recognized, while suppressing variability introduced by speakers, environments, and transfer chains. To be efficient, the features should match the assumptions made during the design of the acoustic model (e.g., the assumption that distributions of the feature vector components can be modeled by limited number of GMM Gaussians).

The majority of features for ASR are derived from the short time spectral envelope of the speech signal. Especially, Mel-Frequency Cepstral Coefficients (MFCC) (Davis and Mermelstein, 1980) and Perceptual Linear Prediction (PLP) (Hermansky, 1990) cepstral coefficients have been popular choices for ASR front-ends recently, Chou and Juan (2003). Block diagrams of MFCC and PLP feature extraction are shown if Fig. 2.2. Some of the extraction stages can be found similar both for MFCC

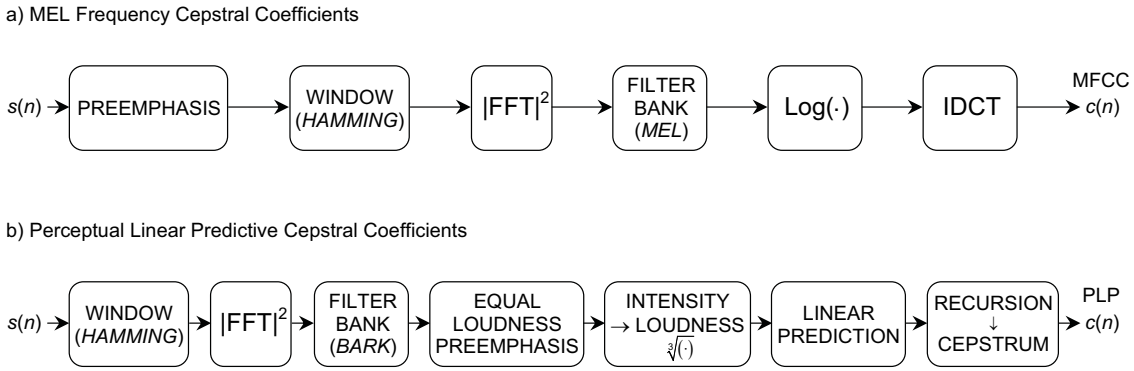

Figure 2.2: Feature extraction of MFCC and PLP cepstral coefficients.

and PLP:

- Signal framing and windowing – the speech signal is divided into overlapping quasi-stationary frames by applying successively shifted Hamming window:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \tag{2.29}$$

  where n is the discrete time and N is the length of the frame/window. In the frequency domain, the windowing is displayed as a convolution of the signal spectrum and the window spectrum. Compared to the rectangular window, spectral side-lobes of the Hamming window are more attenuated, participating less in the distortion of the estimated short-time spectrum of the signal (Harris, 1978). Window length is chosen typically 20–35 ms and the shift 5–15 ms. [2]

- Estimation of the energy spectrum by applying a short-time Fourier transform (STFT) to the windowed signal, Rabiner and Juang (1993):

$$|X(k)|^2 = \left|\mathcal{F}\left\{s(n)\,w(n)\right\}\right|^2 = \left|\sum_{n=m}^{m+N-1} s(n)w(n-m)\,e^{-j2\pi nk/N}\right|^2, \tag{2.30}$$

[2]Sometimes, a pitch synchronous framing is used. Here, the window length and overlap are varied according to the length of the actual pitch period (in the case of voiced speech intervals) or averaged nearest pitch periods (in the case of unvoiced speech intervals).

where $\mathcal{F}$ denotes a discrete Fourier transform, m is the discrete time of the frame beginning, N is the frame length, and k stands for the discrete frequency, $k = 0, 1, \ldots, N-1$.

- Warping and decimation of the energy spectrum by a bank of filters. In MFCC, triangular filters are placed equidistantly on the warped frequency axis called a mel scale (Volkmann *et al.*, 1937):

$$mel\,(f) = 1127 \ln\left(1 + \frac{f}{700}\right). \tag{2.31}$$

  In PLP, trapezoid-shaped filters are distributed equidistantly on the Bark scale (Zwicker, 1961):

$$bark\,(f) = 6 \ln\left[\frac{f}{600} + \sqrt{\left(\frac{f}{600}\right)^2 + 1}\right]. \tag{2.32}$$

  In both cases, the warping and filtering is conducted to simulate the nonlinear frequency resolution and critical-band integration as observed in human hearing. In addition, the trapezoid filters in PLP approximate the shape of the cochlea's critical band masking curves.

- Pre-emphasis – in the first stage of the MFCC extraction, a first-order high-pass filtering is conducted:

$$s_p\,(n) = s\,(n) - \alpha \cdot s\,(n-1)\,, \tag{2.33}$$

  where α is a pre-emphasis factor typically chosen to range in the interval $0.9 \leqslant \alpha < 1$. The purpose of MFCC pre-emphasis is to compensate for the spectral tilt due to modeling of speech using volume velocity versus microphone measurement using sound pressure, which is typically -6 dB/oct (Childers and Lee, 1991). In PLP, an equal-loudness pre-emphasis is applied to the filter bank outputs to simulate varying sensitivity of human hearing across frequencies. Up to 5 kHz, the pre-emphasis weighting function is defined:

$$E\,(f) = \frac{\left(f^2 + 1.44 \cdot 10^6\right) f^4}{\left(f^2 + 1.6 \cdot 10^5\right)^2 \left(f^2 + 9.61 \cdot 10^6\right)}. \tag{2.34}$$

  If a speech signal of wider bandwidth is processed, (2.34) can be further extended by the term representing a decrease of the sensitivity of human hearing occurring above 5 kHz (Hermansky, 1990). Although the pre-emphasis in PLP is primarily motivated by the auditory model of human hearing, in fact, it performs a similar spectral tilt compensation as seen in MFCC (Hönig *et al.*, 2005). The pre-emphasis assures that all frequency components of the spectral envelope will be given an equivalent attention in the subsequent steps of the feature extraction process.

In PLP, the equal loudness pre-emphasis is followed by a cubic-root amplitude compression representing the relation between intensity and loudness. The energy spectrum is then approximated by means of linear prediction, yielding coefficients of an all-pole filter:

$$|X_{LP}\,(k)|^2 = \frac{G^2}{\left|1 - \sum_{l=1}^{p} a_l e^{-jkl}\right|^2}, \tag{2.35}$$

where p is an order of the linear predictor filter, $\{a_l\}$ are the prediction coefficients, and G is the gain constant derived from the mean squared prediction error. The prediction coefficients are derived from the autocorrelation coefficients by applying the Levinson-Durbin recursive algorithm, Rabiner and

Juang (1993). In the last step, cepstral coefficients are extracted. Cepstrum of the energy spectrum envelope is defined, Proakis and Manolakis (1995):

$$c(n) = \mathcal{F}^{-1}\left\{\ln|X_{LP}(k)|^2\right\}. \tag{2.36}$$

Intuitively, coefficients of the energy cepstrum reflect spectral relations in the energy spectrum envelope $|X_{LP}(k)|^2$. Low cepstral coefficients relate to the slow changes in the envelope while the higher coefficients to the faster envelope changes. Specially, $c(0)$ refers to the logarithm of DC component of the energy spectrum, (i.e., to the energy of the signal). The following 1–2 coefficients are typically related to the overall spectral tilt, (Paul, 1987). Considering the speech production mechanism[3], lower cepstral coefficients are driven by the vocal tract shape, while the higher coefficients relate to the vocal tract excitation. Since changes in vocal tract shape are believed to carry the main portion of linguistic information[4] (McCandless, 1974), while the excitation is strongly speaker-dependent and often redundant, it is common to extract only the first 12–14 cepstral coefficients for speech recognition. In PLP, the cepstral coefficients are derived from the predictor coefficients using a recursion algorithm equivalent to performing an inverse Fourier transform of the log spectrum as defined in Eq. (2.36), Rabiner and Juang (1993). Since only the low cepstral coefficients are of interest, a low order LP ($8 \leqslant p \leqslant 14$) approximating slow changes in the energy spectral envelope is used. In MFCC, the cepstral coefficients are calculated from the inverse discrete cosine transform (IDCT) of the logarithm of the filter bank outputs:

$$c(n) = \sum_{q=1}^{Q} \ln|X_q|^2 \cos\left[n\left(q-\frac{1}{2}\right)\frac{\pi}{Q}\right], \tag{2.37}$$

where Q is number of filters in the filter bank, and $|X_q|^2$ is the energy output of the q-th filter.

The MFCC or PLP coefficients $c(n)$ are called static coefficients. It was observed that adding first and second order time derivatives of the static coefficients into the feature set (i.e., providing information about the temporal evolution of the static features) improves performance of the ASR systems, (Furui, 1986), (Hanson and Applebaum, 1990b). The first order time derivatives, delta coefficients, can be obtained from the regression analysis of the time evolution of the static coefficients, (Furui, 1986):

$$\Delta c_i(m) = \frac{\sum_{l=1}^{L} l\left[c_{i+l}(m) - c_{i-l}(m)\right]}{2\sum_{l=1}^{L} l^2}, \tag{2.38}$$

where $\Delta c_i(m)$ is the m-th cepstral coefficient in the feature vector of the i-th frame, and L is a number of frame pairs participating in the linear regression. The second time derivatives, called acceleration or delta-delta coefficients, can be calculated from the linear regression of delta coefficients:

$$\Delta\Delta c_i(m) = \frac{\sum_{l=1}^{L} l\left[\Delta c_{i+l}(m) - \Delta c_{i-l}(m)\right]}{2\sum_{l=1}^{L} l^2}. \tag{2.39}$$

The delta and acceleration coefficients can be implemented also as the first and second order differences of the static coefficients, (Hanson and Applebaum, 1990b):

$$\delta c_i(m) = c_{i+L}(m) - c_{i-L}(m), \tag{2.40}$$

[3]Mechanism of speech production will be discussed in more detail in Sec. 3.1.

[4]In tonal languages, e.g., in Mandarin Chinese, also tonal patterns are employed to distinguish words, see (Lei *et al.*, 2006).

$$\delta\delta c_i\,(m) = \delta c_{i+LL}\,(m) - \delta c_{i-LL}\,(m)\,, \tag{2.41}$$

where L and LL are optional parameters referring to the half of the distance between the frames differenced.

While MFCC and PLP as described here represent mainstream speech coding, several studies have shown that altering the extraction stages can further improve efficiency of the features in ASR systems (e.g., replacing FFT by LPC in MFCC, (Bou-Ghazale and Hansen, 2000), or substituting triangular filters for trapezoid filters in PLP, (Psutka *et al.*, 2001))[5]. In state-of-the-art systems, the basic feature extraction scheme is usually extended by variety of segmental and suprasegmental, fixed and adaptive operations, with focus on increasing the class discriminability and invariance to adverse conditions provided by the features. Some of these techniques are discussed in the following chapters.

2.6 Summary

Putting together parts discussed in the previous sections, the architecture of an HMM-based ASR system may look as shown in Fig. 2.1. In the 'feature extraction' stage, the speech signal is segmented into overlapping frames. From each frame, a feature vector (parameterized acoustic observation) is extracted. The feature vectors are passed to the 'sub-word likelihoods' section, where the likelihoods that the feature vector is generated by the given sub-word model are estimated for all sub-word models using GMM or MLP-based statistical models. In MLPs, to each of the network outputs a single phone probability is usually assigned. In the case of GMMs, the sub-word unit likelihoods are usually modeled by multiple GMM models and transition probabilities, forming sub-word HMMs.

In the final stage, the 'acoustic-phonetic-language' model is employed. In this model, the sub-word models are connected into word models comprising single or multiple pronunciations as defined by the lexicon. The transition probabilities between the word models are given by the language model. Based on the sub-word likelihoods, a path through the 'acoustic-phonetic-language' model giving the maximum likelihood is searched by the Viterbi algorithm. States passed by the maximum likelihood path generate the recognizer's output.

[5] An impact of altering FFT and LPC, as well as filter bank modifications are studied in Chap. 9.

Chapter 3

Lombard Effect (LE): An Overview

In the first decade of the twentieth century, French oto-rhino-laryngologist Etienne Lombard presented results of a series of experiments in which patients were exposed to noise while being engaged in a conversation[1]. Lombard observed that the patients increased their vocal effort and pitch when exposed to noise, and lowered the voice and pitch to the former level once the noise stopped. Similar increase of vocal level was observed with attenuation of the speaker's auditory feedback. If speaker with normal hearing had noise fed into one ear (monaural noise), they raised their voice just slightly as they still could hear themselves with the other ear. If presented to binaural noise, they raised their voice close to shouting. Patients with unilateral deafness reacted only slightly or not at all when having monaural noise fed to the impaired ear, but when the noise was fed to the healthy ear, they started to nearly shout, as their auditory feedback was masked, (Lane and Tranel, 1971).

When reporting the experiments, Lombard noted, that speakers' changes in voice production due to noise seemed to be unconscious. This gave rise to a theory that speech production is kind of automatic servomechanism being controlled by auditory feedback. The theory was supported by results of several experiments, e.g., in (Pick *et al.*, 1989), speakers were unable to follow instructions to maintain constant vocal intensity across alternating periods of quiet and noise. In another experiment of the same work, the speakers learned to suppress consciously the effect of noise by using a visual feedback. However, after the feedback was removed, they tended to lower their overall vocal level both in noise and in quiet, rather than changing their specific response to the noise.

In contrast to this, in other studies, significant differences were observed in speech production when the speakers were communicating (Webster and Klumpp, 1962) or just reading texts (Dreher and O'Neill, 1957), showing that the reaction to noise cannot be purely automatic, but rather consciously driven by the speaker's effort to maintain effective communication. (Lane and Tranel, 1971) suggests that the response to noise may be initially learned through the public loop (loop speaker-listener), and later becomes a highly practiced reaction when communicating in noise.

In the recent study (Junqua *et al.*, 1998), speakers were exposed to noise while communicating with a voice-controlled dialing system. The system was trained in quiet, reaching the best performance when processing neutral speech. The speakers were able to consciously compensate for the Lombard effect and lower their voices to reach efficient response from the system. This confirms the hypothesis stated in (Lane and Tranel, 1971) that the changes in speech production are driven (at least to some extent) by the public loop. These observations lead to the definition of Lombard effect, (Junqua, 1993), (Womack and Hansen, 1999), which can be stated as follows: Lombard effect is the process when speakers change their speech production in an effort to maintain intelligible communication in a noisy environment.

[1]The results were published in (Lombard, 1911). Particular Lombard's experiments and findings are also summarized and further discussed in (Lane and Tranel, 1971).

Lombard's findings significantly contributed to the following areas:

- Tests of hearing loss and malingering,
- Study of dynamics between hearing and speaking,
- Analysis of speech communication in noise.

Inspired by Lombard's experiments, an extensive number of studies have analyzed the impact of LE on speech production. The conditions used to induce LE, as well as the analysis techniques, varied across the studies, hence the resulting observations might contradict in particular cases. Nevertheless, majority of the analyses reported consistent shifts of several speech features. Not only voice intensity and pitch varied under LE as observed by Lombard, but also other excitation and vocal tract parameters were affected.

Recent ASR systems usually rely on the acoustic models trained on the neutral speech uttered in a calm environment – 'clean neutral speech'. Hence, feature shifts introduced by LE and by the presence of additive noise in the speech signal may cause a strong mismatch with the acoustic models, resulting in a severe deterioration of the recognition performance.

The additive noise is present, to some extent, in all speech signals. Its amount may easily reach a level impacting the ASR accuracy, even if the environmental noise itself is not strong enough to induce LE. Hence, alleviating the impact of noise has been a crucial issue on the way to robust recognition. During the last three decades, a variety of efficient noise suppression/speech emphasis algorithms as well as noise-modeling algorithms were proposed.

Even though the impact of LE on speech production were analyzed before the rise of ASR, a relatively little attention has been paid to LE in the design of recognition systems. A necessity to suppress LE emerges with the voice-controlled applications (navigation, telephony, automated information centers, etc.) operated in the real-world environments (e.g., public transport or crowded places). In general, LE corrupts recognition performance considerably even if the noise is suppressed or not present in the speech signal (Junqua *et al.*, 1998), (Hansen, 1996), (Bou-Ghazale and Hansen, 2000), and the performance deterioration by LE can be significantly stronger than the one caused by the corresponding additive noise (Rajasekaran *et al.*, 1986), (Hansen, 1988), (Takizawa and Hamada, 1990). Hence, to improve ASR efficiency in various noisy environments, it is necessary to study feature variations caused by LE, and employ this knowledge in the system design.

The remainder of this chapter presents an overview of the speech production changes under LE, and approaches to Lombard speech modeling, classification, recognition, and acquisition.

3.1 Model of Speech Production

The speech waveform can be modeled as a response of the vocal tract to the series of quasi-periodic glottal pulses or to noise, depending on whether the actual sound is voiced or unvoiced, respectively. The vocal tract acts as a non-uniform tube of time-varying cross-section (Flanagan, 1957). Resonances of the vocal tract are called formants. Frequencies and bandwidths of formants are determined by the shape of the vocal tract. In the natural speech, the configuration of the vocal tract is continuously varied by articulators (tongue, jaw, lips, and velum) to produce distinct speech units.

The vocal tract is terminated by the glottis (the opening between vocal folds) on one side, and by lips on the other side. In voiced speech, the stream of air expelled from the lungs is periodically interrupted by the vibrating vocal folds, producing glottal pulses. The pulses form a quasi-periodic volume velocity waveform. The frequency of the glottal pulses, called fundamental frequency, determines the speech intonation. In unvoiced speech, the stream of air from the lungs passes the glottis without

interruption, as the vocal folds are left open and do not vibrate. In this case, all audible components of speech are produced by the articulators.

A linear digital model of speech production, Rabiner and Schafer (1978), is shown in Fig. 3.1. For voiced speech, an impulse train generator $I(z)$ produces a sequence of unit impulses. Distance between the impulses is driven by the period length of the desired pitch. The impulse sequence excites a glottal pulse model $G(z)$, whose impulse response has a shape of the desired glottal pulse. In modal voice, the average slope of the glottal pulse spectra is approximately -12 dB/oct, (Childers and Lee, 1991). The vocal intensity of the voiced speech is adjusted by the gain A_V. For unvoiced speech, the vocal tract is excited by random noise with a gain A_N. The transfer function of the vocal tract is

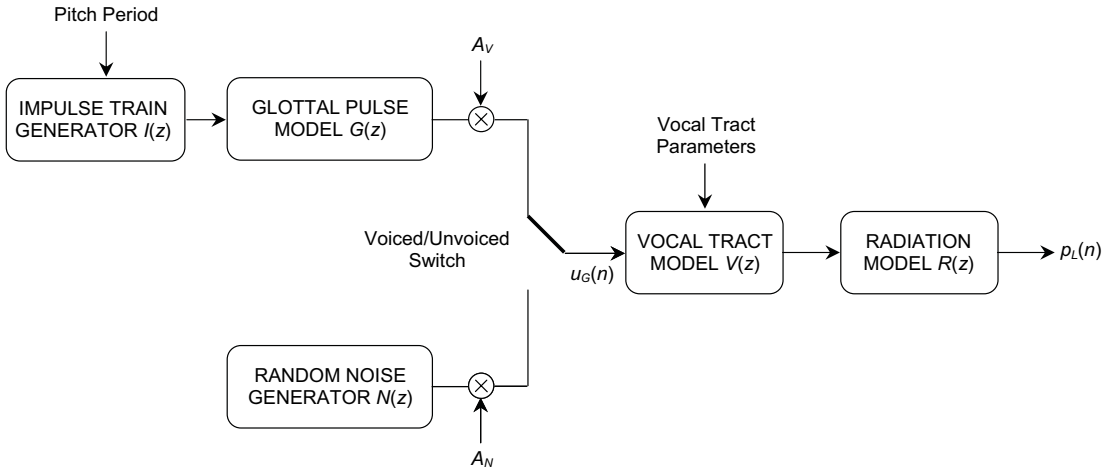

Figure 3.1: Digital model of speech production.

modeled using an all-pole digital circuit[2]:

$$V(z) = \frac{G}{1 - \sum_{k=1}^{N} \alpha_k z^{-k}}, \tag{3.1}$$

where G is the gain constant, N is the order of the filter, $\{\alpha_k\}$ are coefficients of the filter, and z is the z-plane. The all-pole model is a good representation of vocal tract configurations for a majority of speech sounds, with the exception of nasals and fricatives, which require the addition of zeros in the transfer function to model anti-resonances.

In nasals, resonances of the nasal tract form poles and the oral tract is a closed branch causing zeros. Also nasalized vowels require zeros. Here, the nasal cavity represents an open side branch, adding extra poles and zeros to the transfer function (McCandless, 1974). In unvoiced fricatives, articulators are placed close together, forming a narrow constriction. Air is forced through the constriction at high velocity, producing a turbulent flow. This flow generates a noise which acts as excitation for the part of the vocal tract anterior to the constriction (i.e., here, the excitation source is located within the vocal tract rather than in the glottis). In this case, the zeros of the transfer function are caused by the poles of the part of the vocal tract posterior to the constriction (Heinz and Stevens, 1961).

[2]In the early works, the transfer function of the vocal tract was modeled as a cascade or parallel connection of the second order resonator circuits (Dunn, 1950). Although successfully simulating formant frequencies and bandwidths, the second order resonators introduced a spectral tilt which did not appear in the transfer function of the real vocal tract. Digital all-pole filters address this drawback of the analog resonators.

For nasals and fricatives, the anti-resonances can be modeled by including zeros or adding more poles into $V(z)$. The latter approach is preferred as it allows for modeling the transfer function by means of linear prediction (see Sec. 2.5).

The transfer function of sound radiation by lips can be considered to have a constant tilt of approximately +6 dB/oct for a wide range of mouth opening areas (Flanagan, 1957). In the digital domain, the lip radiation can be modeled as a first order high-pass filter:

$$R(z) = R_0(1 - z). \tag{3.2}$$

The complete digital model of speech production can be then expressed:

$$S(z) = \begin{cases} A_V I\,(z)\, G\,(z)\, V\,(z)\, R\,(z)\,, & \textit{voiced}, \\ A_N N\,(z)\, V\,(z)\, R\,(z)\,, & \textit{unvoiced}. \end{cases} \tag{3.3}$$

Considering the average spectral tilt −12 dB/oct of the glottal pulses in modal speech and the constant tilt of +6 dB/oct introduced by lips, the average tilt of the modal voiced speech spectra reaches −6 dB/oct. For a majority of speech sounds, it can be assumed that the parameters of the excitation and vocal tract are stationary in 10–30 ms time segments. This assumption is often used as a cue to speech signal segmentation in feature extraction.

In some modes of speech production, the presented model is rather a crude approximation of reality and has to be further refined. For example, in breathy speech, the glottis has an imperfect closure and, during the vocal folds vibration, the vocal tract is excited both by glottal pulses and by noise, (Childers and Lee, 1991). In such a case, the switch in Fig. 3.1 should be replaced by a summation unit. Another assumption made in the present linear model is that the airflow propagates through the vocal tract as a plane wave. However, it was shown in (Teager, 1980) that the airflow rather creates vortices near the walls of the vocal tract, resulting in separation of the flow into several components, see Chap. 2–3 in Jr. *et al.* (2000). Based on this observation, a Teager energy operator (TEO) was proposed and successfully employed in the speech style classification, (Hansen and Womack, 1996) see Sec. 3.4.

3.2 Speech Production under LE

Parameters of neutral speech can be analyzed relatively easily as extensive multi-lingual corpora have been acquired in the past decades. On the other hand, occurrence of LE in the corpora is quite rare. The first step in the works focusing on LE analysis typically comprises acquisition of Lombard speech in actual or simulated noisy conditions. In the majority of past studies, utterances from only a limited number of subjects were considered and the results obtained could hardly be generalized to a larger population. However, many of the observations are consistent across works, outlining a speaker-independent picture of the impact of LE on speech production. The following sections attempt to summarize these observations, with a focus on feature variations affecting ASR or providing a cue to Lombard speech classification.

3.2.1 Vocal Intensity and Speech Intelligibility

To express vocal intensity, sound pressure level (SPL) is commonly used:

$$L_p = 20 \log_{10} \left(\frac{p_{RMS}}{p_0} \right), \tag{3.4}$$

where p_{RMS} is the root mean square (RMS) sound pressure and p_0 is a reference pressure, $p_0 = 20\mu$Pa RMS for air environment. The logarithmic expression of the sound pressure correlates almost linearly with the perceived loudness of sound, (Fletcher and Munson, 1933).

Lombard noticed that speakers engaged in conversation tend to raise their *vocal intensity* when exposed to noise[3], (Lane and Tranel, 1971). Later, a relation between noise level and corresponding voice level was studied. It was found that for a wide range of noise levels, the dependency between voice SPL and noise SPL, called *Lombard compensation function*, is almost linear. In (Dreher and O'Neill, 1957), for subjects reading a text in noise, Lombard function had a slope 0.1 dB/dB. In another study, (Webster and Klumpp, 1962), pairs of subjects communicating in various levels of ambient noise had a direct spoken feedback and were required to maintain relatively error-free communication. In this case, Lombard function had a slope 0.5 dB/dB. Presumably, the slope is steeper here because speakers were motivated to react to noise in a way to be understood by others[4]. Similar slope of Lombard function was reported in (Lane *et al.*, 1961) also for altered sidetone gain and speech level. When the sidetone level was increased, speakers lowered their voices, but only halfway compared to the sidetone increase. Lombard function reaches its minimum in low noises (below 50 dB SPL) where speakers do not further decrease their vocal intensity with the decrease of noise. In very high levels of noise (above 106 dB SPL) Lombard function approaches maximum as speakers reach limits of capability to further increase their vocal level, (Lane *et al.*, 1970).

The increase of vocal effort is not uniform across phones produced in noise. Vowels seem to be more emphasized than consonants, causing a decrease of *consonant-vowel energy ratio*, (Hansen, 1988), (Junqua, 1993), (Womack and Hansen, 1996a).

Furthermore, *intelligibility* of speech produced in noise was evaluated in perceptual tests. In (Dreher and O'Neill, 1957), speakers read utterances in several noise conditions and in quiet. The noise was delivered to their headsets, yielding speech recordings with high SNR. For the perceptual tests, noise was added to the recordings to reach a constant SNR. Utterances originally produced in noise were found to be more intelligible than utterances produced in quiet. A similar observation was made in (Pittman and Wiley, 2001). Female Lombard speech was found more intelligible in noise than male Lombard speech, (Junqua, 1993). The higher the level of noise in which the speech was originally produced, the higher the intelligibility in perceptual tests was reached in the constant SNR, (Summers *et al.*, 1988). However, when the vocal effort is increased after a certain point, the intelligibility of speech starts to deteriorate, (Pickett, 1956), (Lane and Tranel, 1971), (Junqua, 1993). In (Pickett, 1956), the decrease of intelligibility was found to be almost linear with the increase of vocal intensity in the range 80–90 dB SPL. Similar deterioration was observed with very weak voice. The rate of deterioration may vary depending on the number of confusable items in the vocabulary to be recognized, (Junqua and Anglade, 1990).

3.2.2 Pitch

In pitch perception theory, the fundamental frequency of glottal pulses F_0 is often called 'spectral pitch' and the perceived pitch is referred to as 'virtual pitch', (Terhardt *et al.*, 1982). In complex sounds (e.g., in vowels) virtual pitch may differ from F_0 as higher harmonics of F_0 are sometimes perceptually more prominent than the fundamental component. It is quite common to use a term 'pitch', which can represent either virtual or spectral pitch, depending on the context (e.g., pitch detection algorithms (PDAs) estimate F_0 or virtual pitch, see Sec. 4.1). In this work, the term pitch will refer to 'virtual' pitch when reporting perceptual observations, and to 'spectral' pitch in most other cases, since the use of F_0-tracking PDAs is prevalent both in the referred works and in the experiments carried out within the thesis. There is strong correlation between spectral pitch and virtual pitch, (Duifhuis *et al.*, 1982), hence, it can be assumed that the observations made for spectral pitch can be generalized also for virtual pitch, and vice versa.

[3]Similar increase of intensity can be found in loud speech, (Hansen, 1988), (Bond and Moore, 1990).

[4]In general, speakers raise their voices proportionally to the subjectively perceived level of the disturbance introduced by noise. The perceived level is speaker-dependent, (Junqua *et al.*, 1999).

Lombard noticed in his experiments that increasing vocal intensity in noise was accompanied by changes in pitch (Lane and Tranel, 1971). Later works found significant changes in pitch contour, mean, variance, and distribution when comparing neutral and Lombard speech, (Hansen, 1988), (Bond *et al.*, 1989), (Junqua, 1993), (Hansen, 1996). An example of typical distribution of pitch in neutral speech and speech uttered in 85 dB SPL pink noise is shown and compared to angry speech in Fig. 3.2, as presented in (Hansen, 1996). In noise, the distribution of pitch may become more Gaussian, (Pisoni *et al.*, 1985). Less significant increase in pitch was observed for female speakers when compared to male speakers, (Junqua and Anglade, 1990).

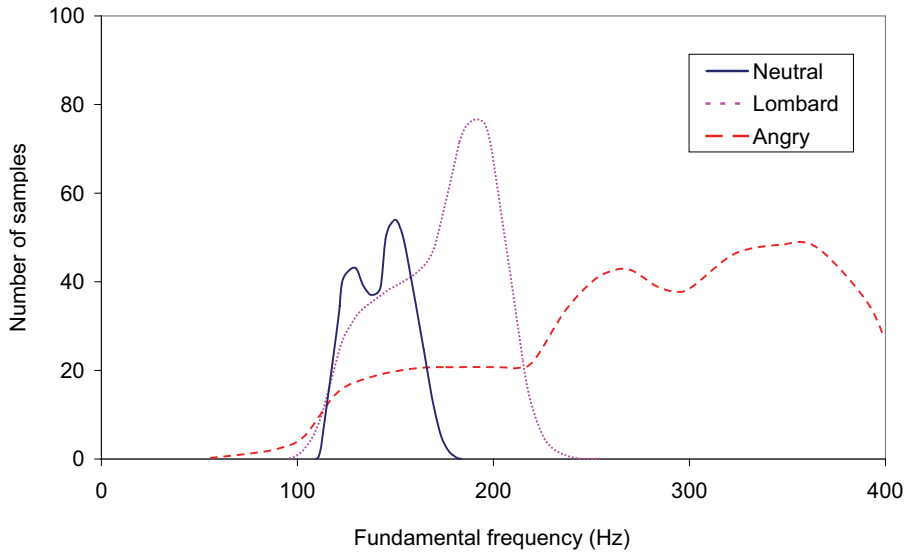

Figure 3.2: F_0 distribution in neutral, LE, and angry speech, after (Hansen, 1996).

Changes of fundamental frequency are physiologically related to the vocal effort. The considerable increase in fundamental frequency of loud and Lombard speech is caused by increased sub-glottal pressure and increased tension in the laryngeal musculature, (Schulman, 1985). Fundamental frequency changes almost linearly with vocal intensity, when expressed in semitones and SPL, respectively, (Gramming *et al.*, 1987), (Titze and Sundberg, 1992), (Sulter and Wit, 1996).

3.2.3 Formants

In voiced speech, frequency of the first formant F_1 varies inversely to the vertical position of the tongue and the second formant F_2 varies with the posterior-anterior articulatory dimensions, (Kent and Read, 1992). Due to wider mouth opening during the speech production, accompanied by lowering the jaw and the tongue, the center frequency of F_1 increases in the voiced loud and Lombard speech, (Schulman, 1985), (Summers *et al.*, 1988), (Bond and Moore, 1990). The increase is independent on the phoneme context, (Junqua and Anglade, 1990). F_2 tends to increase in some phones, (Junqua, 1993), while may decrease in others, (Bond *et al.*, 1989), see also Fig. 3.3. In (Pisoni *et al.*, 1985), the increases of F_1 frequency were accompanied by consistent decreases of F_2 while in (Hansen, 1988), (Hansen and Bria, 1990) locations of both $F_{1,2}$ increased for most phonemes. In (Takizawa and Hamada, 1990), formants occurring bellow 1.5 kHz tended to shift upwards in frequency while the higher formants shifted downwards. The higher formants, the smaller degree of resultant shift was observed. Average bandwidths of the first four formants decrease for most phonemes, (Hansen, 1988), (Hansen and Bria, 1990), (Junqua, 1993).

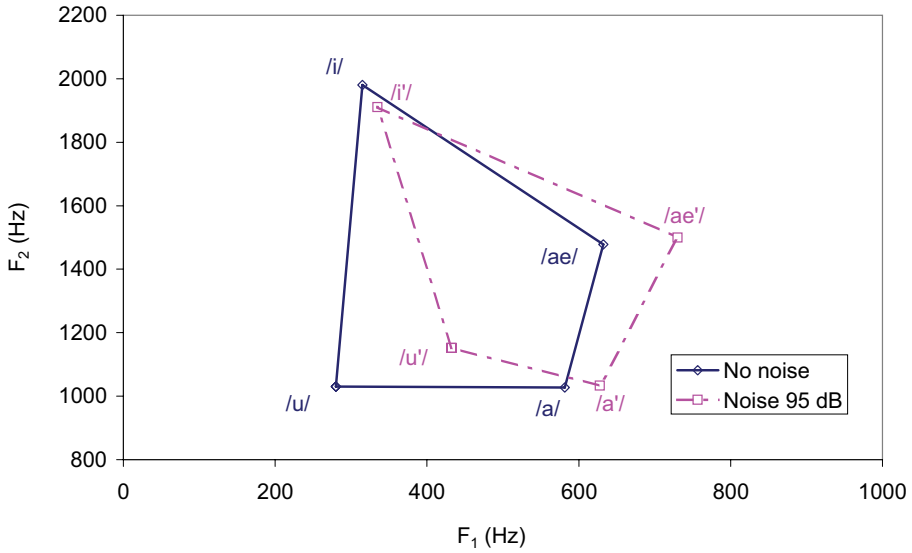

Figure 3.3: Average center frequencies of F_1 and F_2 in neutral and LE vowels, after (Bond et al.*, 1989).*

3.2.4 Spectral Slope

As discussed in Sec. 3.1, the speech waveform is generated by joint contribution of vocal tract excitation, vocal tract transfer function, and radiation by lips. The radiation by lips has an almost constant spectral tilt of +6 dB/oct for wide range of mouth opening areas, (Flanagan, 1957), and its contribution to spectral slope changes in short-time speech segments can be neglected.

The vocal tract can be modeled by a lossless tube with zero average spectral slope, Rabiner and Schafer (1978). It can be assumed that the slope would not change under LE, as only the first two formants were observed to shift significantly in Lombard speech, see Sec. 3.2.3, while higher formants remained almost intact. However, due to the band-limiting of speech signals for the purposes of digital processing, the former vocal tract spectrum comprising infinite number of formants is typically reduced to just the first 3–5 formants. In this case, the redistribution of $F_{1,2}$ due to LE may affect the estimated spectral slope. Presumably, the contribution will be quite limited, since the shifts of $F_{1,2}$, even if consistent and statistically significant, reach negligible values compared to the whole speech bandwidth from which the spectral slope is extracted, see, e.g., (Junqua and Anglade, 1990). The main contribution to the spectral changes can be credited to the variations of glottal pulse spectra. In (Monsen and Engebretson, 1977), significant variations of shape and intensity spectrum of glottal volume-velocity waveforms were found when analyzing speech uttered in various talking styles, including neutral and loud speech. Similar observations were made in (Cummings and Clements, 1990), where glottal waveforms estimated by inverse filtering of speech samples from eleven talking styles, including LE, were found to have unique time-domain profiles. In the frequency domain, the overall spectral slope decreases in loud and Lombard speech[5], (Pisoni *et al.*, 1985), (Summers *et al.*, 1988), (Hansen, 1988), (Hansen and Bria, 1990), accompanied by upward shift of spectral center of gravity, (Junqua, 1993). In loud and LE speech, speakers tend to concentrate energy into the frequency range of the highest sensitivity of the human auditory system. This energy migration happens at the expense of low and high frequency bands. As a result, low-band spectral slope (0–3 kHz) becomes more gradual and high-band spectral slope (3–8 kHz) becomes steeper, (Stanton *et al.*, 1988). Similar observations were also made for linguistically stressed syllables in neutral speech, (Sluijter and van Heuven, 1996), (Crosswhite, 2003).

[5]The spectral slope of loud and Lombard speech spectral may reach similar values, (Hansen, 1996).

3.2.5 Duration and Speech Rate

When producing speech in noise, syllables are often prolonged, (Dreher and O'Neill, 1957). Duration of vowels tends to increase while duration of consonants is reduced. The duration decrease in consonants is usually smaller than the duration increase in vowels, resulting in an increase of average word durations, (Junqua and Anglade, 1990), (Lane and Tranel, 1971). Word duration changes were found either significant, (Hansen, 1988), (Hansen, 1996), or insignificant, (Bond *et al.*, 1989), depending on the actual conditions.

As already discussed in Sec. 3.2.1, besides the increase in duration, vowels are also produced with an increased vocal effort compared to consonants. This may be caused by the fact that vowels are the most audible speech sounds at high noises and long distances, carrying a major part of the information to the listener, and are therefore intentionally emphasized by the speaker, (Junqua, 1993).

Speech rate is closely related to the duration of speech units. Some studies observed a decrease of speech rate under LE, (Dreher and O'Neill, 1957), (Webster and Klumpp, 1962). In the latter work, the speech rate decreased with the increasing noise but tended to increase with increasing number of listeners (to a certain point). Others have not found any correlation between LE and speech rate, (Lane and Tranel, 1971).

3.3 Degradation Model of LE

As discussed in the previous sections, LE affects a wide range of speech production parameters. To propose ASR algorithms more resistant to LE, it is useful to first consider a general model of speech degradation by LE, that is, a model transforming clean neutral speech into noisy Lombard speech. In state-of-the-art systems, speech recognition features are extracted from short-time spectra, hence, it is reasonable to search a degradation model representing short-time spectral changes introduced by noisy environment. In (Chi and Oh, 1996), the proposed degradation model has the following form:

$$S_{LE}(\omega) = G \cdot A(\omega) S_N \left[F(\omega) \right] + N(\omega), \tag{3.5}$$

where $S_N[\cdot]$ is the spectrum of clean neutral speech, $F(\omega)$ is the nonlinear frequency warping representing variations of formant locations and bandwidths, $A(\omega)$ is the amplitude scaling related to the redistribution of energy between frequency bands and changes of spectral tilt, G is the gain constant representing vocal intensity variations, and $N(\omega)$ is the spectrum of additive noise. Furthermore, the model (3.5) can be extended for environmental convolutional distortion, and convolutional and additive distortions introduced by the transfer chain, (Hansen, 1988), (Hansen, 1996).

Parameters of the degradation model can be found by confronting clean neutral and corresponding noisy Lombard speech samples (Chi and Oh, 1996). The degradation model can be employed both in adapting ASR systems to LE, or in 'equalizing' Lombard speech towards neutral, which would be better accepted by neutral-trained systems. Another way to model LE is to use source generator model, (Hansen, 1994). Here, the differences between neutral and LE speech are represented by a set of transformations mapping neutral speech production features (e.g., vocal tract, duration, intensity, glottal source factors) to LE ones.

3.4 Classification of Neutral/LE Speech

In general, talking style classification finds application in various areas, such as in the development of human-machine interfaces responding and adapting to the user's behavior, (Lee and Narayanan, 2005), redirecting highly emotional calls to a priority operator, or monitoring aircraft communication (Womack and Hansen, 1996a).

Neutral/LE speech classification discussed in this section is intended to be used for improving recognition performance in changing noisy conditions. As mentioned earlier, the mismatch between LE speech features and neutral-trained acoustic models may result in severe deterioration of the recognition performance. The mismatch can be reduced by using a neutral/LE classifier to divert the speech to be recognized into dedicated acoustic models trained on neutral and Lombard speech, (Womack and Hansen, 1996b), see also Sec. 3.5.

When designing a classifier, it is crucial to find a set of features providing sufficient discriminability between classes to be recognized. Many of the speech features discussed in the previous sections display significant shifts between neutral and LE conditions. However, distributions of some of them may be strongly context or speaker-dependent, and thus would not provide a reliable cue to the classification. The following features were found to be efficient in actual neutral/LE classification experiments:

- Normalized energy, (Bou-Ghazale and Hansen, 1998),
- Glottal pulse shape, (Cummings and Clements, 1990), and spectral tilt, (Zhou *et al.*, 1998), (Hansen, 1996),
- Pitch mean, (Kwon *et al.*, 2003), normalized pitch mean, (Bou-Ghazale and Hansen, 1998), (Zhou *et al.*, 1998), variance, (Hansen, 1996), delta and acceleration, (Kwon *et al.*, 2003),
- Phone class duration, (Womack and Hansen, 1996a), vowel class duration, (Zhou *et al.*, 1998), intensity of voiced sections, (Womack and Hansen, 1996a), (Zhou *et al.*, 1998),
- Energy shifts from consonants toward vowels, (Womack and Hansen, 1996a).
- Estimated vocal tract area profiles and acoustic tube coefficients, (Hansen and Womack, 1996),
- Mel-band energies, mel-frequency cepstral coefficients (MFCC), (Kwon *et al.*, 2003), autocorrelation MFCC, (Hansen and Womack, 1996),
- Linear predictive cepstral coefficients, (Bou-Ghazale and Hansen, 1998),
- Teager energy operator (TEO)[6], (Hansen and Womack, 1996), (Zhou *et al.*, 1998).

Mean vowel $F_{1,2}$ locations are not suitable for classification, (Zhou *et al.*, 1998). This finding is intuitively supported by the fact that locations of $F_{1,2}$ vary significantly within each of the talking styles due to articulation of distinct speech sounds, see Fig. 3.3. MFCC delta and acceleration coefficients also do not perform well, (Womack and Hansen, 1996a), for being resistant to talking style changes, (Lippmann *et al.*, 1987), (Hanson and Applebaum, 1990a).

Talking style classifiers typically employ GMMs, (Zhou *et al.*, 1998), (Neiberg *et al.*, 2006), HMMs, (Zhou *et al.*, 1998), k-nearest neighbors (KNN)[7], (Lee and Narayanan, 2005), artificial neural networks (ANNs), (Womack and Hansen, 1996a), and support vector machines (SVMs)[8], (Kwon *et al.*, 2003).

Apart from the choice of classification feature set, the classification performance also depends on the length of the speech segment being analyzed (represented by the feature vector). It was observed that with increasing length of the feature vector the classification error rate tends to drop (Zhou *et al.*, 1998).

[6]Instead of propagating uniformly through the vocal tract as a plane wave, the airflow creates vortices due to the energy concentrated near the walls. TEO measures energy resulting from this nonlinear process, (Hansen and Womack, 1996)

[7]Class-labeled training samples form vectors in a multidimensional feature space. Distances between the classified sample and all training samples are calculated. The classified sample is assigned to the class which is the most common among its k nearest neighbors, (Cover and Hart, 1967).

[8]Support vector machines map vectors to a higher dimensional space. In the space, a hyperplane splitting samples into two classes is constructed. The hyperplane is searched to maximize the distance between its hypersurface and the nearest training samples of both classes, (Burges, 1998).

3.5 Robust Speech Recognition

The performance of recent ASR systems is strongly impacted by environmental and speaker-related factors. The environmental factors comprise additive noises and convolutional distortions due to reverberation and microphone/telephone channels. The speaker-related factors are represented by differences in speech production across speakers (physiological differences, dialects, foreign accents), and by conscious and reflexive changes in speech production on the speaker level (talking styles, emotions, Lombard effect, task load, etc.). The physiological properties of speech production correlate with age and gender of speakers, (Junqua, 2002), being displayed in differences in glottal pulse shapes and frequencies, and shapes and lengths of the vocal tract[9].

In particular, the performance of ASR systems may significantly deteriorate due to the impact of noise, (Mokbel and Chollet, 1995), reverberation, (Junqua, 2002), inter-speaker variability, (Womack and Hansen, 1996a), and variability on the speaker level, (Rajasekaran *et al.*, 1986), (Bou-Ghazale and Hansen, 1998). A variety of algorithms improving ASR robustness to these factors has been proposed, operating on the level of model training, front-end processing, and back-end processing. Since the occurrence of LE is displayed in the speaker-related variations of speech features and in the speech signal contamination by noise, many of the methods popular in the 'general' robust ASR design are, to a certain extent, efficient also when dealing with Lombard speech. In the *model training*, the following approaches were found efficient:

- Multi-style training – training models on speech comprising various talking styles has shown to improve performance of the speaker-dependent system as the training and decoding algorithms focused attention on spectral/temporal regions being consistent across styles, (Lippmann *et al.*, 1987), (Chen, 1987). However, in the speaker-independent system, the multi-style training resulted in worse performance than the training on neutral speech, (Womack and Hansen, 1995a).

- Style-dependent/speaker-independent training – models were trained and tested with speech from the same talking style, reaching improvements over neutral trained recognizer both for training on actual or synthetic stressed speech, (Bou-Ghazale and Hansen, 1998), see Fig. 3.4. In the stressed speech synthesizer, variations of pitch contour, voiced duration, and spectral contour were modeled for angry, loud, and Lombard speech. The speaker-independent models were trained with the differences (perturbations) in speech parameters from neutral to each of the stressed conditions. In the synthesis stage, the trained perturbation models were used to statistically generate perturbation vectors to modify the talking style of input neutral speech. In (Iida *et al.*, 2000), a concatenative speech synthesizer[10] was used to produce emotional speech (joy, anger, sadness).

- Training/testing in noise – models were trained and tested in the same noisy conditions, (Yao *et al.*, 2004).

Improved training methods can increase recognition performance for the conditions captured in the training sets, but the performance tends to degrade when the conditions change, (Hansen, 1996).

In the *front-end processing* stage, the aim is to provide speech representation preserving linguistic message carried in the speech signal, while being invariant to environmental/speaker-induced changes. A number of feature extraction and feature equalization techniques has been proposed. The feature extraction techniques employ:

- Auditory-based models, (Seneff, 1986), (Ghitza, 1988), (Hermansky, 1990), (Mak *et al.*, 2004),

[9] Vocal tract length varies typically from 18 cm for males to 13 cm for females and children, and to 7 cm for new-born babies (Lee and Rose, 1996), (Vorperian *et al.*, 2005).

[10] Concatenative speech synthesizer strings together units of recorded natural speech.

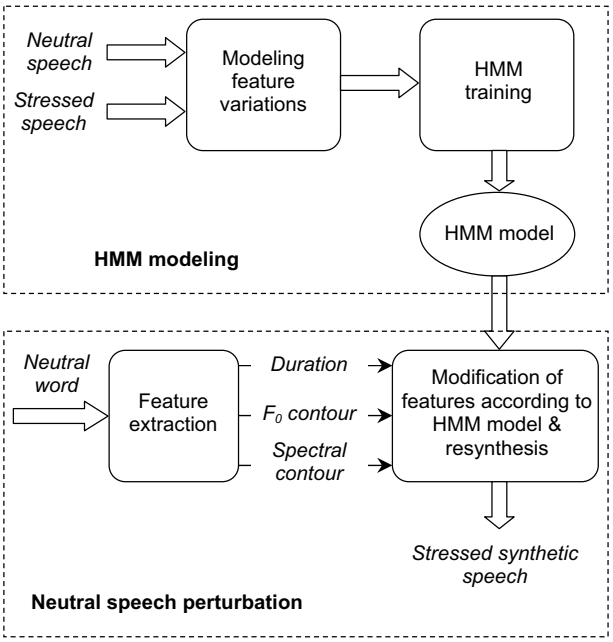

Figure 3.4: HMM-based stressed speech synthesis, after (Bou-Ghazale and Hansen, 1998).

- Temporal-spectral transformations – FFT and LPC (Davis and Mermelstein, 1980), (Hermansky, 1990), wavelets, (Gallardo *et al.*, 2003), Minimum Variance Distortionless Response (MVDR) – power spectrum derived from the output of data-dependent bandpass filters, (Yapanel and Hansen, 2008),
- Optimized filter banks, (Biem and Katagiri, 1997), (Bou-Ghazale and Hansen, 2000), (Gallardo *et al.*, 2003),
- Temporal information – delta, second (acceleration), and third derivatives of feature vectors, (Lippmann *et al.*, 1987), (Hanson and Applebaum, 1990a).

The feature equalization techniques employ:

- Noise suppression – linear (Boll, 1979), (Hansen and Bria, 1990), and nonlinear (Lockwood and Boudy, 1991) spectral subtraction of noise spectrum, speech enhancement, (Hansen and Clements, 1991),
- Cepstral mean subtraction – fixed, (Lee and Rose, 1996), and adaptive, (Bou-Ghazale and Hansen, 2000),
- Cepstral variance normalization, (Paul, 1987),
- Spectral tilt compensation, (Stanton *et al.*, 1989), (Hansen and Bria, 1990), (Takizawa and Hamada, 1990), (Suzuki *et al.*, 1994), (Bou-Ghazale and Hansen, 2000),
- Formant shifting/vocal tract length normalization (VTLN), (Hansen and Clements, 1989), (Takizawa and Hamada, 1990), (Suzuki *et al.*, 1994), (Lee and Rose, 1996),
- Formant bandwidth compensation, (Takizawa and Hamada, 1990), (Suzuki *et al.*, 1994),
- Duration compensation, (Takizawa and Hamada, 1990),

- Whole-word cepstral compensation, (Chen, 1987), (Hansen, 1988), (Hansen, 1996),
- Segmental cepstral compensation, (Wakao *et al.*, 1996)
- Source generator based adaptive cepstral compensation, (Hansen, 1996).

In the *back-end processing*, the following approaches were found efficient:

- Adding noise characteristics into clean reference models, (Mokbel and Chollet, 1995), parallel model combination (PMC), (Gales and Young, 1996),
- Model adaptation to actual speaker and conditions – maximum likelihood linear regression (MLLR), (Gales and Woodland, 1996), maximum a posteriori estimation (MAP), (Gauvain and Lee, 1994),
- Building stronger duration models – dying exponential e^{-at} modeling state duration likelihoods in standard HMMs is replaced by the peak-shaped function te^{-at} which better represents average state durations, (Paul, 1987),
- Signal decomposition to speech and noise components using 2–D HMM – the dimensions contain speech and noise models respectively, a 2–D state sequence tracing speech and noise models yielding maximum likelihood is searched by extended Viterbi algorithm, (Varga and Moore, 1990),
- Talking style decomposition by N–channel HMM – to each talking style one HMM channel is assigned. As the style changes, the style-dependent models yielding maximum likelihood are traced, (Womack and Hansen, 1999), see Fig. 3.5.

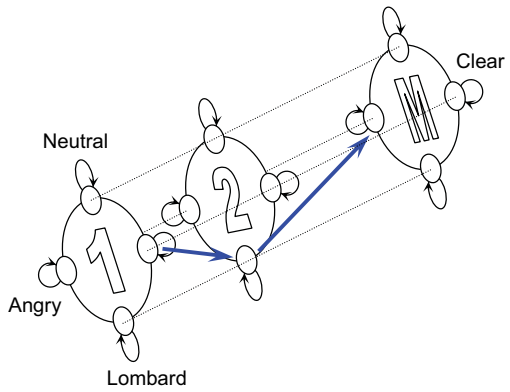

Figure 3.5: N-channel HMM, after (Womack and Hansen, 1999).

- Talking style decomposition by stress classifier followed by codebook of dedicated recognizers, (Womack and Hansen, 1996b), or weighting the output of codebook of dedicated recognizers by stress classifier, (Womack and Hansen, 1995b).

Selected methods from the areas of front-end and back-end processing are further discussed in the remainder of the thesis.

3.6 Lombard Speech Corpora

Currently, extensive multi-lingual corpora covering a variety of environments attractive for the application of human-machine interfaces are available. The environmental scenarios range from calm places (e.g., offices, providing clean neutral speech recordings, to adverse noisy places, such as public places and cars, (Iskra *et al.*, 2002), (CZKCC, 2004), capturing noisy speech recordings). These databases are very valuable for training and testing ASR systems intended to operate in the real-word conditions. On the other hand, they are not always the best choice for the analysis of LE. The most significant drawbacks of the data recorded in the real adverse conditions are:

- Changing level and spectral properties of background noise,
- Strong speech signal contamination by noise,
- Absence of communication factor in the recordings – typically, subjects just read text without need to preserve intelligibility of the speech in noisy conditions, (Junqua *et al.*, 1998).

These factors make it difficult to study separately the impacts of noise and LE on ASR, analyze reliably variations of speech feature distributions (performance of feature trackers tend to decrease with noise), or find a relationship between actual noise level and speech feature shifts (the actual noise levels are not known).

For this reason, the majority of works analyzing LE have preferred to collect Lombard speech data in simulated noisy conditions[11], where noise was introduced to the speaker's ears through calibrated headphones while the speech was sensed by a close-talk microphone. This setup yields recordings with high SNR. Within the literature, the following background noises were used in simulated noisy conditions:

- Pink noise: 65 dB SPL, 75 dB SPL, (Varadarajan and Hansen, 2006), 85 dB SPL, (Hansen and Bou-Ghazale, 1997), (Varadarajan and Hansen, 2006), 90 dB SPL, (Stanton *et al.*, 1988), (Suzuki *et al.*, 1994), 95 dB SPL, (Rajasekaran *et al.*, 1986), (Chen, 1987), (Bond *et al.*, 1989),
- Highway noise: 70 dB SPL, 80 dB SPL, 90 dB SPL, (Varadarajan and Hansen, 2006),
- Speech shaped noise: 40–90 dB SPL, (Korn, 1954), 85 dB SPL, (Lippmann *et al.*, 1987),
- Large crowd noise: 70 dB SPL, 80 dB SPL, 90 dB SPL, (Varadarajan and Hansen, 2006),
- White Gaussian noise: 80 dB SPL, (Wakao *et al.*, 1996), 85 dB SPL, (Hanson and Applebaum, 1990b), (Junqua and Anglade, 1990), 80–100 dB SPL, (Summers *et al.*, 1988),
- Various band-limited noises: 80 dB SPL, (Wakao *et al.*, 1996).

In several works, speakers were provided with adjustable speech feedback (sidetone) allowing for compensation of the sound attenuation caused by wearing headphones. In (Bond *et al.*, 1989), speakers adjusted their sidetone to a comfortable level in quiet in the beginning of the session, since then the level was kept constant both for recordings in quiet and noisy conditions. In (Junqua *et al.*, 1998), the sidetone level adjustments were allowed throughout the whole session. This scheme might have blurred differences between the speech produced in quiet and the speech produced in noise as in the

[11] Advantages and disadvantages of using simulated noisy conditions can be found similar to those mentioned in (Murray and Arnott, 1993) for elicited emotional speech: "A trade-off exists between realism and measurement accuracy of emotions generated by speakers in the laboratory (questionable realism, but verbal content and recording conditions controllable) and field recordings (real emotions, but content and recording conditions less controllable)."

noisy conditions, speakers tended to compensate for the sidetone masking by increasing the sidetone level instead of just increasing their vocal effort.

The databases of simulated Lombard speech successfully address the first two drawbacks of the real adverse data, i.e., speech is uttered in noises of defined spectrum and intensity, and the speech signal reaches high SNR. However, the third factor, the need to maintain intelligible communication in noise, is often missing in the recordings. Speakers typically just read text without having feedback whether their speech is intelligible to others, (Junqua, 1993). In this case, speakers increase their vocal effort rather due to the masking of their auditory feedback by noise (physiological effect), than due to the need to reach an efficient communication, (Junqua *et al.*, 1998).

The intelligibility factor was employed in the simulated noisy scenario in (Korn, 1954), where a speaker and an operator were communicating while wearing headphones fed by noise of the same level. In (Webster and Klumpp, 1962), pairs of subjects were communicating in actual noise, with the requirement that the listener repeat each word received. If the repetition was incorrect, the speaker said the same word again. Such scenarios assured that speakers would maintain intelligibility of their speech. In the recent work, (Junqua *et al.*, 1998), speakers were communicating with an automatic dialing system while listening to noise through headphones. Since the dialing system was trained for neutral speech, in spite of listening to noise, speakers had to produce speech close to 'neutral' to communicate efficiently with the system. The results from this study are very valuable, proving that speakers consciously modify their speech production when communicating. On the other hand, this scenario is not a good example of typical communication in noise.

The Lombard speech databases were usually acquired for the purposes of the particular studies, comprising utterances from just a few speakers, and were not made publicly available[12]. Two exceptions are the SUSAS database[13] (Speech under Simulated and Actual Stress), (Hansen and Bou-Ghazale, 1997), and UT-Scope database (Speech under Cognitive and Physical stress and Emotion), (Varadarajan and Hansen, 2006). SUSAS provides a comprehensive collection of English utterances produced in various conditions. The database comprises a variety of talking styles, emotional speech, speech uttered while performing computer tasks, and speech uttered in simulated noisy conditions. UT-Scope contains speech produced under cognitive and physical stress, emotional speech (angry, fear, anxiety, frustration), and Lombard speech. The Lombard speech recordings were collected in 3 simulated background noise scenarios. In each scenario, speakers were exposed to 3 levels of noise ranging from 65 to 90 dB SPL.

Considering the Czech language, several large databases recorded in real conditions are available. In the presented thesis, Czech SPEECON[14], (Iskra *et al.*, 2002), and CZKCC, (CZKCC, 2004), databases were employed in the experiments. Czech SPEECON comprises speech recordings from homes, offices, public places, and cars. CZKCC contains recordings from the car environment. The SPEECON office sessions provide samples of neutral speech with high SNR. The SPEECON and CZKCC car recordings were promising to contain LE, because the speech was uttered in increased levels of noise there. However, the initial analyses of the SPEECON and CZKCC car data have shown a very limited presence of LE, see Chap. 6. Hence, to obtain data suitable for LE experiments, a new database of simulated Czech Lombard speech was acquired. The design, recording setup, and contents of the database are discussed in the following chapter.

[12]Note that the fact that research groups do not usually release their corpora makes it difficult to duplicate results of past studies on LE.

[13]SUSAS database is available through the Linguistic Data Consortium, (LDC, 2008).

[14]SPEECON database is distributed through the European Language Resources Association, (ELRA, 2008).

Chapter 4

Experimental Framework

This chapter describes tools used in the presented work for feature analyses and speech recognition. Methods for analyzing pitch, formants, vocal intensity, spectral slope, duration, and also evaluation measures and the recognition setup are discussed.

For the purposes of pitch and formant tracking, a variety of publicly available tools is available. In particular, WaveSurfer, (Sjolander and Beskow, 2000), and Praat, (Boersma and Weenink, 2006) were employed in the analyses. There are several pitch tracking algorithms implemented in Praat. To decide which tool and algorithm to use, the Praat and WaveSurfer algorithms, as well as a novel pitch detection algorithm (PDA) developed by the author were compared on the reference pitch-labeled database.

4.1 Pitch

As already noted in Sec. 3.2.2, pitch can be viewed from two different sides. It can represent either the fundamental frequency of glottal pulses (F_0) – 'spectral pitch', or the sound pitch as perceived by the human auditory system – 'virtual pitch'. In complex sounds, virtual pitch may differ from spectral pitch since the spectral component perceived as the strongest one does not need to be the F_0 component, (Terhardt, 1974). A vast number of spectral and virtual pitch PDAs have been designed during the last decades. Pitch correlates with prosodic features such as lexical stress, tone quality, and sentence intonation, (Seneff, 1986), and represents a strong cue to the talking style assessment, see Sec. 3.2.2. PDAs play an important role in pitch synchronous speech analysis and synthesis, such as in pitch-synchronous feature extraction, triggering the glottal pulses in vocoders, or driving the pitch-synchronous overlap-and-add (OLA) synthesis. The fact that virtual pitch PDAs were also successfully employed in the speech synthesis systems, (Duifhuis *et al.*, 1982), shows that there is a strong correlation between spectral and virtual pitch.

The algorithms for *spectral pitch* tracking typically employ the following approaches:

- Time domain waveform: Speech signal filtering to reduce contribution of higher formants on the waveform, extraction of features from the waveform (amplitudes and distances of peaks, valleys, zero-crossings), combining the features to estimate F_0, e.g., (Gold and Rabiner, 1969),

- Autocorrelation (AC), cross-correlation (CC): The F_0 candidate is determined from the distance between maxima in the AC, (Boersma, 1993), or CC, (Acero and Droppo, 1998), of the segmented speech signal. Autocorrelation weighted by the inverse of the average magnitude difference function (AMDF), (Shimamura and Kobayashi, 2001), or autocorrelation of the LPC residual, (Secrest and Doddington, 1983), were also found effective for finding F_0.

- Cepstral analysis: Low cepstral coefficients of voiced speech represent vocal tract/glottal waveform shapes while the higher components relate to pitch. Pitch period is determined by searching a strong peak in the cepstrum starting from the coefficient related to the highest expected pitch, (Schafer and Rabiner, 1970), (Ahmadi and Spanias, 1999).
- Amplitude spectrum (DFT, wavelets): The F_0 candidate is chosen from the maxima in the amplitude spectrum, (Janer, 1995),
- Joint time and frequency domain: Combination of time and frequency domain techniques, e.g., (Liu and Lin, 2001), (Janer, 1995).

From the domain of spectral tracking, Praat autocorrelation (Praat_ac) and cross-correlation (Praat_cc) algorithms, and WaveSurfer cross-correlation algorithm were chosen to participate in the performance test. In the AC/CC based methods, speech signal is first windowed. In *Praat_ac*, (Boersma, 1993), the unbiased estimation of the original segment AC is obtained by dividing the AC of the windowed signal by the AC of the weighting window. This eliminates the tilt present in the AC of the windowed signal and simplifies the subsequent thresholding and peak picking. In *Praat_cc*, CC between two speech segments is calculated. The candidate pitch period is determined from the window shift yielding maximum in CC function. In *WaveSurfer*, Talkin's RAPT CC algorithm is implemented, (Talkin, 1995). Here, after performing CC and finding pitch candidates from peak picking, dynamic programming is performed to decide the resulting pitch estimate sequence.

The algorithms for *virtual pitch* tracking are typically based on the virtual pitch theory or on using harmonic sieve, or a combination of both methods. The virtual pitch theory, (Terhardt *et al.*, 1982), assumes that each spectral component of the voiced sound generates a series of sub-harmonics. These sub-harmonics are combined in the central pitch processor of the auditory system, following the phenomena of frequency masking and spectral dominance, resulting in the perceived pitch.

In the harmonic sieve approach, (Duifhuis *et al.*, 1982), the central pitch processor attempts to match a harmonic pattern to the analyzed speech spectrum. The harmonic pattern comprises the candidate pitch and its weighted higher harmonics. The matching is realized by passing the speech signal spectrum through the harmonic sieve (i.e., summing narrow regions in the speech spectrum corresponding to the harmonics of the candidate pitch).

From the domain of virtual pitch tracking, sub-harmonic summation method (SHS) and spatial pitch network (SPINET) are implemented in Praat. In the SHS algorithm (Praat_shs), (Hermes, 1988), a spectral compression is performed by using a logarithmic frequency scale. On the linear frequency scale, harmonic peaks of F_0 appear at $F_0, 2F_0, 3F_0, \ldots$, while on the logarithmic frequency scale they occur at $\log(F_0), \log(F_0) + \log(2), \log(F_0) + \log(3), \ldots$ Hence, the distances between harmonics of the compressed spectrum are independent on F_0. To find the pitch estimate, spectral energies of the harmonics are summed for each F_0 candidate. The F_0 yielding the maximum sum is chosen as the resulting pitch:

$$Pitch = \arg\max_{F_0} \sum_{k=1}^{N} \left| X\left[\log(F_0) + \log(k)\right]\right|. \tag{4.1}$$

This approach follows the concept that each spectral component activates not only the element of the central pitch processor that is most sensitive to this component, but also the elements most sensitive to its sub-harmonics.

In SPINET, (Cohen *et al.*, 1995), the speech segment is first passed through the bank of 512 bandpass filters placed in the region 50–5000 Hz, modeling the basilar membrane filtering. The output of the bank is further filtered by the bandpass representing the transfer of the outer and middle ear. In the next stage, cooperative interactions across the nearby frequencies and competitive interactions across a broader frequency band are evaluated using an artificial neural network. Finally, the harmonic summation is carried out.

4.1.1 Design of Novel Time-Domain Pitch Tracking Algorithm

A novel algorithm for direct time domain fundamental frequency estimation (DTFE) and voiced/unvoiced (V/UV) classification is presented in this section. The goal is to design an algorithm for the real-time pitch detection, providing time and frequency resolution comparable to AC-based algorithms while significantly reducing the computational costs. The DTFE algorithm comprises spectral shaping, adaptive thresholding, and F_0 candidate picking based on applying consistency criteria[1]. A scheme of the DTFE algorithm is shown in Fig. 4.1.

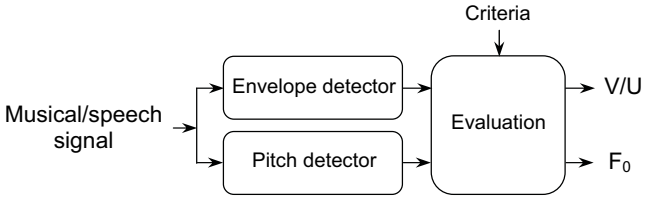

Figure 4.1: DFE chain.

Envelope Detector

The envelope is determined as a short-time moving average of the signal energy realized by an FIR filter. Since all coefficients of the moving average filter are identical ($1/M$, where M is the filter order), instead of weighting and summing all actual values present in the filter buffer, only the input signal sample is weighted and inserted to the buffer while the last buffer sample is removed (LILO – last in, last out). The length of the moving average window (0.025–0.027 sec) is set as a compromise between envelope smoothing and ability to follow fast energy changes on the boundaries of voiced/unvoiced parts of the speech signal.

Pitch Detector

Higher harmonic components of the glottal pulse spectrum are emphasized by the vocal tract resonances and tend to form multiple peaks in the single pitch period of the time domain speech waveform. This makes it difficult to estimate length of the pitch period directly from the distance between adjacent peaks in the original waveform. To suppress these additional peaks, the speech signal is low-pass filtered (spectral shaping). The low-pass filter is designed to introduce spectral tilt starting at 80 Hz, assuring that, starting from the second harmonics of F_0, all spectral components will be suppressed even for low F_0 values[2]. To minimize transient distortion induced by fast amplitude changes of the filtered signal, low order IIR filter (3^{rd} order Butterworth) is used for spectral shaping. An example of spectral shaping by the low-pass filter is shown in Fig. 4.2. The thin dashed line represents spectrum of the original signal, the bold dashed line depicts transfer function of the low-pass filter, and the solid line plots the spectrum of the filtered signal. Corresponding time domain signals are shown in Fig. 4.3. After the spectral shaping, all local extremes are detected. Due to the low order of the low-pass filter, some 'false' peaks and zero-crossings may still remain in the signal. To identify locations of the significant extremes, adaptive significant peak picking based on neighboring

[1]The DTFE was originally developed as a monophonic musical pitch detector for the guitar MIDI converter, (Bořil, 2003a), (Bořil, 2003b), and later adapted for the pitch tracking in speech signals, (Bořil and Pollák, 2004).

[2]For modal voices, the typical occurrence of F_0 can be expected approximately in 80–600 Hz. For vocal fry, F_0 was observed to reach frequencies 24–90 Hz, and F_0 for falsetto reached 280–630 Hz in males and 500–1130 Hz in females, (Monsen and Engebretson, 1977), (Childers and Lee, 1991).

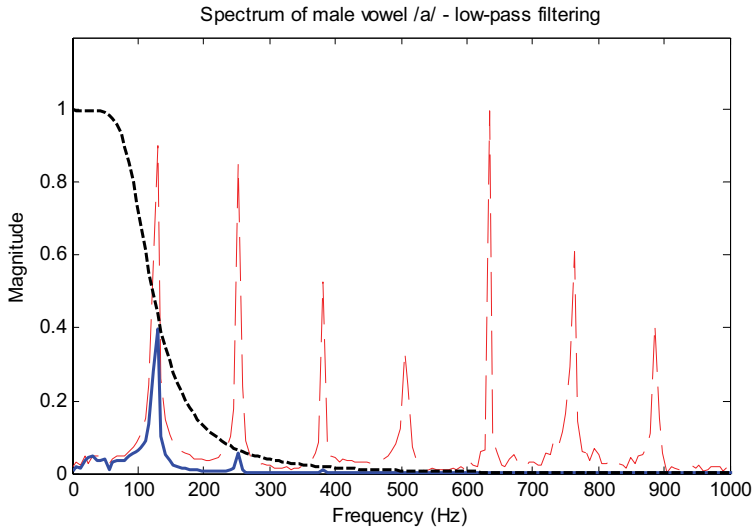

Figure 4.2: Spectral shaping by low-pass IIR.

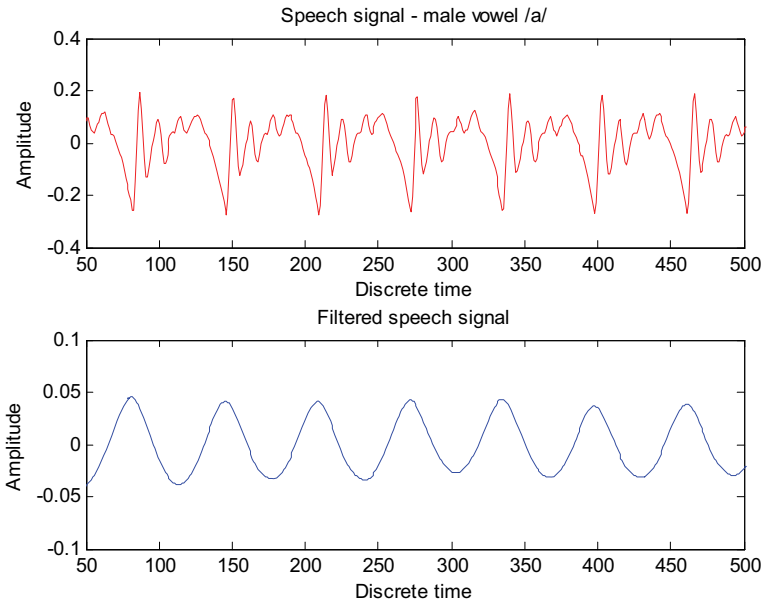

Figure 4.3: Speech signal – male vowel /a/ before and after spectral shaping.

peaks thresholding is performed. P_1 is termed *a significant peak* related to the maximum if:

$$P_1 > 0 \ \cap \ ZC(P_{last}, P_1) = 1 \ \cap \ P_1 > P_2 \cdot th \ \cap \ [P_1 > P_2 \ \cup \ ZC(P_1, P_2) = 1]\,, \tag{4.2}$$

where $ZC(X, Y) = 1$ if there is at least one zero-crossing between peaks X and Y, else 0. P_{last} denotes the previous significant peak. In the subsequent iteration, P_2 is shifted to P_1, a new peak becomes P_2 and the test is repeated. The significant peak related to the minimum is obtained by reversing the signs of the inequality in Eq. (4.2). Finally, the pitch period estimate is determined from the distance between adjacent significant peaks related to the maxima or minima respectively. Examples of adaptive peak picking are schematically shown in Fig. 4.4. Robustness of significant peak picking

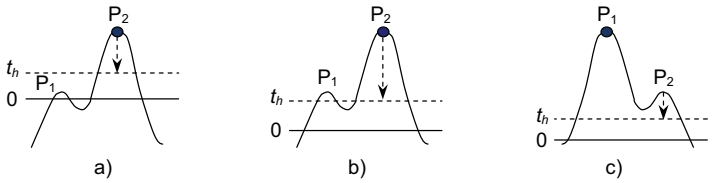

Figure 4.4: Adaptive peak picking.

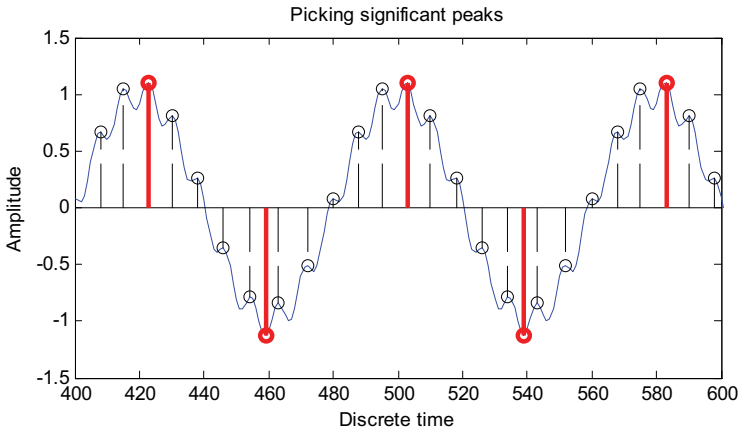

Figure 4.5: Example of adaptive peak picking in signal corrupted by additive harmonic noise.

to quasi-stationary additive noise is demonstrated in Fig. 4.5. The complete pitch detector flow diagram is shown in Fig. 4.6.

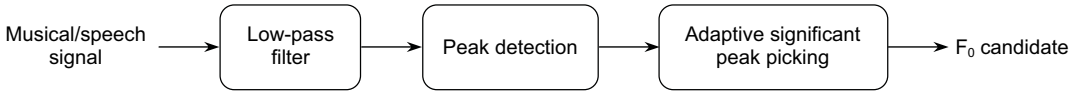

Figure 4.6: Pitch detector flow diagram.

Evaluation Unit

The pitch detector passes all F_0 candidates to the evaluation unit, including estimates from the unvoiced segments and segments of speech silence. Moreover, in the voiced speech candidates, energy doubling and halving appears frequently. To eliminate these misleading pitch cues, several consistency criteria are applied. The first criterion defines an energy threshold E_{th}. No estimates are accepted for the signal levels lower than E_{th}:

$$E_k < E_{th} \;\Rightarrow\; F_{est} \neq F_k. \tag{4.3}$$

The actual energy E_k is obtained from the envelope detector. The second criterion defines a range of accepted F_0 values (typically set to 80–600 Hz):

$$F_k \notin (F_{floor}; F_{ceiling}) \;\Rightarrow\; F_{est} \neq F_k. \tag{4.4}$$

As the third criterion, a newly proposed M–order majority criterion is used. The M–order majority criterion requires that more than half of M consecutive estimates lie in the same frequency band of

the defined width. The criterion represents a non-linear filtering of the discrete sequence of estimates. Similarly as in the case of median filtering, (Tukey, 1974), the 'noisy' samples deviating from the general trend are eliminated. However, unlike the case of median filtering, continuity and rate of trend variation are also controlled here by the order of the criterion and width of the band. The order of the criterion determines a rate of smoothing the resulting F_0 estimates. The higher the value of M, the better the immunity to frequency doubling and halving. On the other hand, increasing M decreases sensitivity of the criterion to the correct estimates in the case of short voiced segments or fast F_0 changes, as increasing number of consecutive estimates are required to lie in the same band. The latter problem can be addressed by widening the frequency band. In the presented experiments, the frequency bandwidth was set to one semitone. Let F_m be M successive candidates and $F_k \in \{F_m\}$. Let $count_{F_k}\{F_m\}$ be a number of F that

$$F \in \{F_m\} \ \cap \ F \in \left(\frac{F_k}{\sqrt[24]{2}}; F_k \cdot \sqrt[24]{2} \right). \tag{4.5}$$

The interval in Eq. (4.5) is set equal to the frequency bandwidth of one semitone centered at F_k. The candidate defining a frequency band covering the maximum of the M estimates is found as:

$$p = \max_k \left(count_{Fk}\{F_m\} \right), \ q = \arg\max_k \left(count_{Fk}\{F_m\} \right), \ k = 1, ..., M. \tag{4.6}$$

If more than half of the M candidates lie in the semitone band with the center frequency F_q, the resulting estimate F_{est} is set equal to F_q:

$$p > \left\lfloor \frac{M}{2} \right\rfloor \ \Rightarrow \ F_{est} = F_q, \tag{4.7}$$

where braces $\lfloor \rfloor$ represent the floor function (round down). If more than one candidate F_k satisfies (4.6) and (4.7), the estimate with the lowest index value is selected (i.e., the first satisfying estimate). An example of the majority criterion being applied to F_0 candidates is shown in Fig. 4.7. In Fig. 4.8, immunity of the majority criterion to the frequency doubling is demonstrated. The dashed line represents the output of the majority criterion unit.

The aforementioned criteria also serve as a V/UV detector. The speech segment which provides F_0 passing all the criteria is labeled as voiced.

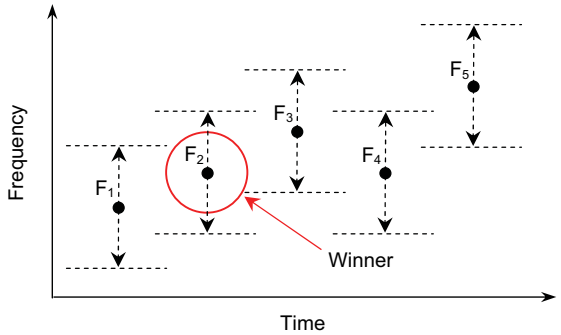

Figure 4.7: 5^{th} order majority criterion – example of F_{est} picking.

Computation Demands

As shown in the previous sections, the novel DTFE algorithm is very computationally efficient. In the envelope detector, the input signal sample is weighted by a constant and fed to the buffer. In

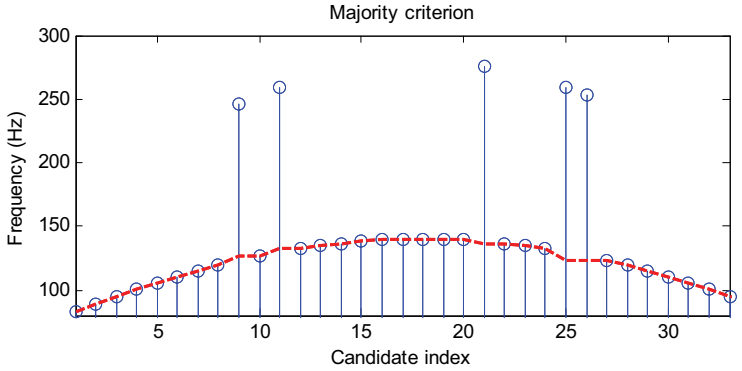

Figure 4.8: 5^{th} order majority criterion – immunity to frequency doubling.

the pitch detector, a 3^{rd} order IIR filter is used for spectral shaping. Here, seven multiplications of variables by a constant, three summations and three subtractions are conducted per signal sample. During peak detection in the pitch detector, the sample value is compared to the previous one. The adaptive peak picking and evaluation is performed only a few times per F_0 period and also employs just multiplications by a constant and comparing amplitudes.

Alternatively, in the case of autocorrelation based methods, hundreds of variable-by-variable multiplications per frame are required, even for low sampling frequencies. Larger computation demands can be found in cross-correlation algorithms, where, unlike autocorrelation based methods, the 'effective overlap' of the correlated segments does not reduce with the window shift.

4.1.2 Evaluation on ECESS Database

The DTFE algorithm and the pitch tracking algorithms available in Praat and WaveSurfer were evaluated on the ECESS[3] PMA/PDA reference database designed for the 1^{st} ECESS PDA/PMA Evaluation Campaign, (Kotnik *et al.*, 2006). The database is a subset of the Spanish SPEECON database, (Iskra *et al.*, 2002):

- Speech signals: 1 hour per channel (4 channels), F_s = 16 kHz, 60 speakers (30 males and 30 females), 1 minute per speaker,
- Content: 17 utterances per speaker – 1 connected digit sequence, 1 money amount, 10 phonetically rich sentences, 5 isolated words,
- Environments: car, office, public places (living rooms, exhibition areas),
- Recording channels – office and public places: close talk (C0), hands free (C1), directional microphone (placed 1 meter from a speaker) (C2), omni-directional microphone (placed 2–3 meters from a speaker) (C3),
- Recording channels – car: close talk (C0), hands free (C1), microphone in the closest front corner of the car cabin (C2), distant front corner microphone (C3).

In the database, the most negative peaks in the speech waveform and V/UV/non-speech segments were manually labeled for the close talk channel. The reference F_0 values were determined from the distances of the adjacent negative peaks. The F_0 values were sampled in 1 ms steps and stored in

[3]European Center of Excellence on Speech Synthesis, (ECESS, 2007).

reference files. Time labels for the distant channels were derived from the close talk channel labels using cross-correlation time alignment.

The following evaluation criteria were used in the 1^{st} ECESS PDA/PMA Evaluation Campaign:

- Voiced (VE) and unvoiced (UE) error: The percentage of voiced (unvoiced) speech segments misclassified as unvoiced (voiced).
- Gross error high (GEH) and gross error low (GEL): The percentage of voiced speech segments where $F_{est} > 1.2 \cdot F_{ref}$ or $F_{est} < 0.8 \cdot F_{ref}$, respectively. F_{est} is the estimated pitch and F_{ref} denotes the reference pitch.
- Absolute difference between the mean values (AbsMeanDiff): Absolute difference between the mean values of the reference and estimated pitch over the whole signal (in Hz).
- Absolute difference between the standard deviations (AbsStdDiff): Absolute difference between the standard deviations of the reference and estimated pitch over the whole signal (in Hz).

Since the mean pitch and its variance are extracted from the whole 1 hour long signal, the aforementioned absolute difference measures are rather coarse, not providing any information about the pitch deviations on the frame level. For this reason, in this thesis, the set of criteria was extended for mean pitch difference (mean difference) and standard deviation of the pitch difference (standard difference). Let mean difference (in semitone cents) be defined:

$$\overline{\Delta}_{\%} = \frac{1200}{N} \cdot \sum_{n=1}^{N} \log_2 \frac{F_{est}(n)}{F_{ref}(n)}, \tag{4.8}$$

where N is the number of compared frames. For example, $\overline{\Delta}_{\%} = 100\,\%$ represents a semitone difference. Let the standard difference (in semitone cents) be defined:

$$\sigma_{\%} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} \left[1200 \log_2 \frac{F_{est}(n)}{F_{ref}(n)} - \overline{\Delta}_{\%} \right]^2}. \tag{4.9}$$

The database channels sensed by four microphones placed at different distances from the speaker allow for evaluation of the PDA performance in various levels of the noisy background. Mean values and standard deviations of channel SNRs are shown in Table 4.1, with channel SNR histograms shown in Fig. 4.9. SNR was estimated using the algorithm discussed in Sec. 4.3.

Parameter	Channel 0	Channel 1	Channel 2	Channel 3
$\overline{\text{SNR}}$ (dB)	27.2	11.1	11.1	4.5
σ_{SNR} (dB)	7.9	9.4	7.5	8.3

Table 4.1: ECESS PMA/PDA reference database – means and standard deviations of channel SNRs.

WaveSurfer cross-correlation algorithm (*WaveSurfer*), Praat autocorrelation (*Praat_ac*), cross-correlation (*Praat_cc*), sub-harmonic summation (*Praat_shs*), and spatial network (*SPINET*) algorithms, and the novel DTFE (*DTFE no fws*) algorithm were evaluated on the reference database, see Table 4.2 and Fig. 4.10, 4.11. From the 'VE+UE' table, it can be seen that both virtual pitch trackers implemented in Praat, *Praat_shs* and *SPINET*, failed to determine voiced and unvoiced speech segments even in the channel with the highest SNR – C0. In the case of *SPINET*, VE contributed

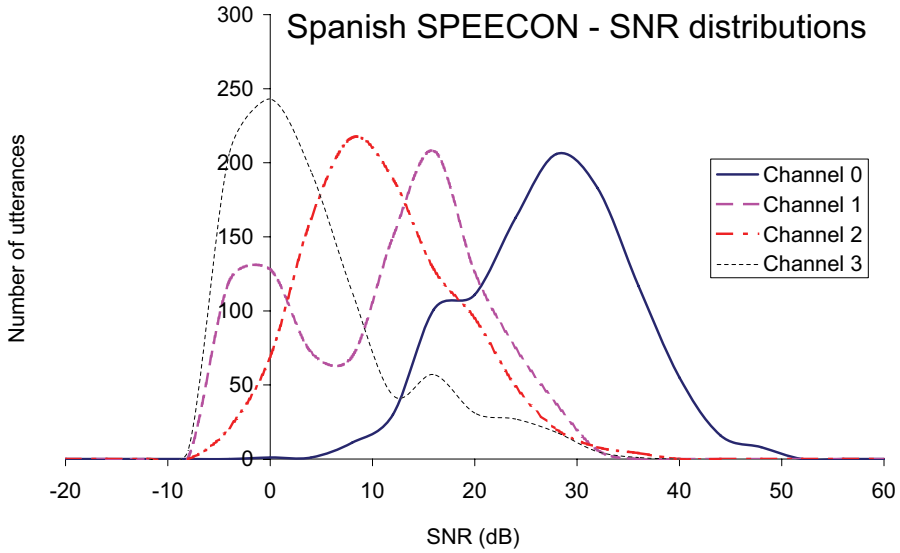

Figure 4.9: ECESS PMA/PDA reference database – channel SNR histograms.

dominantly to the overall error (VE = 88.30 %, UE = 0.08 %) (i.e., the algorithm was almost insensitive to speech). In the case of *Praat_shs*, the disproportion between VE and UE was considerably lower (VE = 17.45 %, UE = 28.07 %), however, the performance was still very poor. Hence, results from *SHS* and *SPINET* are not discussed further.

Across the remaining algorithms, the 'VE+UE' error was comparable on C0, with *WaveSurfer* showing the best performance and *'DTFE no fws'* performing slightly worse than *Praat_cc*. Considering the rest of the evaluation parameters, all spectral pitch trackers reached comparable values on the C0 channel. Comparing channels C0–C4, the 'VE+UE' error grows rapidly for all algorithms with increasing microphone distance from speakers, as the signal waveform starts to be severely distorted by the additive noise and convolutional noise. The error increase is most significant in the case of *'DTFE no fws'*, autocorrelation and cross-correlation algorithms display better resistance to noise.

To improve the noise immunity of *'DTFE no fws'*, the algorithm was extended for noise suppression (NS) based on spectral subtraction[4], yielding a setup denoted *'DTFE'* in the tables and figures. NS was conducted in the input stage of the pitch detector. To avoid the occurrence of musical noise typical for NS employing half-wave rectification, full-wave rectification algorithm from the open source tool CTUCopy, (Fousek, 2007), was used. As shown in Table 4.2, the *'DTFE'* algorithm reaches comparable accuracy of voiced signal segmentation as *WaveSurfer*, *Praat_ac*, and *Praat_cc* on the channel C1, and outperforms *Praat_ac* and *Praat_cc* on the channels C2–C3. *'DTFE'* provides significantly better accuracy of pitch tracking on the channel C0 compared to *WaveSurfer*, *Praat_ac*, and *Praat_cc*, and on the channel C1 compared to *Praat_ac*, and *Praat_cc*, see $\overline{\Delta}_{\%}$ and $\sigma_{\%}$ parameters.

Compared to the other algorithms, *WaveSurfer* displayed the best results in the majority of parameters across the analyzed channels. Possibility to further improve its performance by adding the same NS as used in *'DTFE'* was explored, see *'WaveSurfer fws'* in Table 4.2. It can be seen that adding the NS helped to eliminate considerably the 'VE+UE' error on the channels C1–C3.

[4]In the single channel spectral subtraction algorithms, voice activity detector (VAD) is usually used to determine the segments of speech silence. From these segments, characteristics of additive noise are estimated. Assuming that the additive noise is quasi-stationary, actual estimate of the noise amplitude spectrum is subtracted from the subsequent speech part of signal. After the subtraction, amplitude spectrum is not guaranteed to be positive. To remove the negative components, half-wave rectification zeroing the negative portions, (Boll, 1979), or full-wave rectification taking absolute values, (Faneuff and Brown, 2003), can be used. Half-wave rectification introduces 'musical' noise artifacts.

VE+UE (%)	DTFE no fws	DTFE	WaveSurfer	WaveSurfer fws	Praat_ac	Praat_cc	Praat_shs	Praat SPINET
C0	14.17	12.99	10.28	11.59	12.88	12.05	45.52	88.38
C1	45.80	32.39	31.76	20.88	31.11	31.60	52.80	95.24
C2	55.43	38.03	31.72	25.17	37.93	39.58	59.51	97.68
C3	73.94	52.52	49.55	33.86	54.01	59.60	67.06	98.81

GEH+GEL (%)	DTFE no fws	DTFE	WaveSurfer	WaveSurfer fws	Praat_ac	Praat_cc	Praat_shs	Praat SPINET
C0	2.41	2.80	2.58	3.03	2.33	3.36	3.17	8.50
C1	4.91	5.16	3.37	4.20	13.17	12.43	5.69	4.05
C2	13.35	13.40	3.44	3.54	6.66	7.46	7.49	2.01
C3	15.42	17.28	5.06	5.57	11.99	14.62	12.49	2.28

$\overline{\Delta}_{\%}$	DTFE no fws	DTFE	WaveSurfer	WaveSurfer fws	Praat_ac	Praat_cc	Praat_shs	Praat SPINET
C0	-13.83	2.64	-20.03	-19.87	-9.22	-19.18	-36.29	-1107.67
C1	-84.31	-35.89	17.36	29.87	250.79	190.91	-54.04	-711.44
C2	-128.94	78.37	-23.37	13.83	28.05	-3.62	0.15	-647.27
C3	-217.52	4.51	-54.41	22.76	106.72	65.58	-104.54	-426.36

$\sigma_{\%}$	DTFE no fws	DTFE	WaveSurfer	WaveSurfer fws	Praat_ac	Praat_cc	Praat_shs	Praat SPINET
C0	188.50	194.66	192.98	208.94	222.45	249.81	240.63	729.65
C1	339.38	315.97	369.39	388.23	837.33	791.29	364.37	745.28
C2	547.06	559.28	296.79	310.95	538.61	510.63	513.87	864.42
C3	673.73	612.72	381.99	368.05	769.58	760.63	556.95	700.59

Table 4.2: Performance of PDAs on ECESS PMA/PDA reference database.

4.1.3 Conclusions

In the evaluation tests on the ECESS PDA/PMA reference database, the newly proposed *DTFE* algorithm displayed comparable performance to WaveSurfer cross-correlation and *Praat autocorrelation* and *cross-correlation* algorithms on the close talk channel, while considerably saving computational costs. When extended for the noise subtraction front end, *DTFE* reached a performance comparable to *Praat autocorrelation* and *cross-correlation* algorithms on all channels. *Praat SHS* and *SPINET* virtual pitch trackers failed to distinguish between voiced and unvoiced segments of the speech signal. The best results across channels were reached by *WaveSurfer*, while adding noise subtraction front end even increased accuracy. For this reason, the *WaveSurfer cross-correlation* algorithm (RAPT) was chosen as the tool for the pitch analyses carried out in the remainder of the thesis.

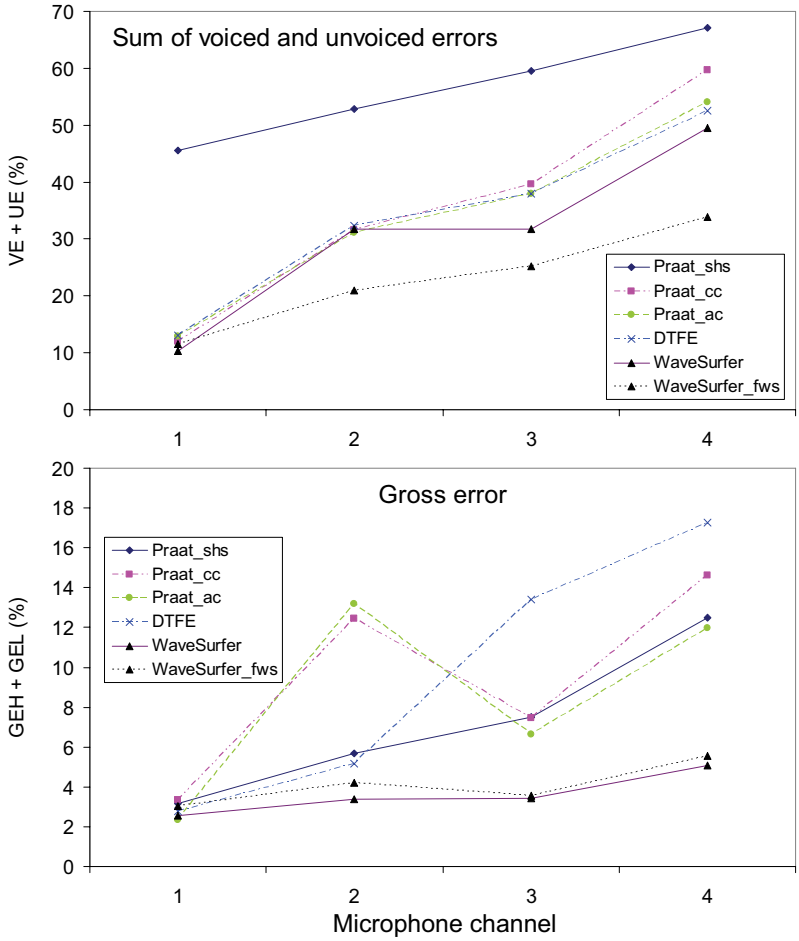

Figure 4.10: Performance of PDAs on ECESS PMA/PDA reference database: $VE + UE$, $GEH + GEL$.

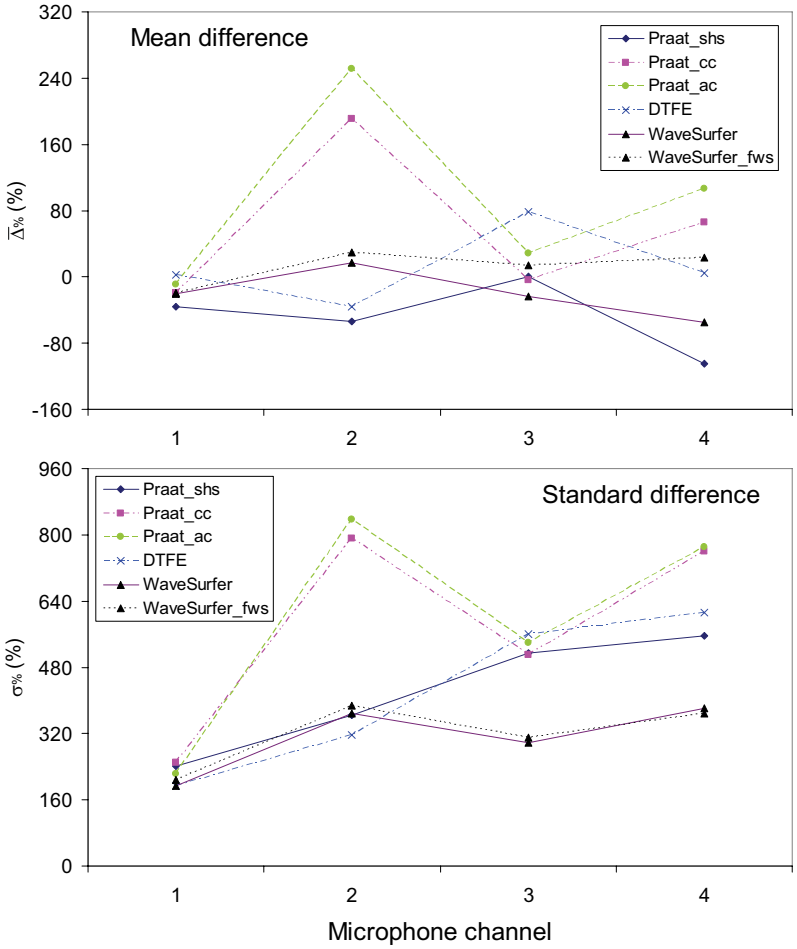

Figure 4.11: Performance of PDAs on ECESS PMA/PDA reference database: $\overline{\Delta}_{\%}$, $\sigma_{\%}$.

4.2 Formants

Formant frequencies play a dominant role in the production of distinct speech sounds. Also, as already mentioned in Sec. 3.2.3, locations and bandwidths of first formants are considerably affected by LE and can provide a cue to talking style assessment. A variety of formant tracking algorithms have been proposed, typically operating in the frequency domain:

- Peak-picking: Finding the location of spectral peaks in the short-time amplitude spectrum, (Flanagan, 1956). The smoothed spectrum can be obtained from cepstrum, (Schafer and Rabiner, 1970) or from the LPC analysis (inverse filtering), (Markel, 1972). In the latter case, the formant candidates are picked from the complex roots of the denominator in the linear predictor,

- Analysis by synthesis: A best spectral match yielding minimum mean square error is synthesized by systematically varying format frequencies and bandwidths in the frequency domain, (Bell *et al.*, 1961), or in the time domain, (Pinson, 1963). In the latter case, a portion of the acoustic waveform is approximated by the systematic variation of amplitudes, phases, damping and oscillation frequencies of a sum of complex exponential functions,

- First order moments: Formant frequencies are calculated as the first order moments within separated portions of the spectrum, (Suzuki *et al.*, 1963).

Formants are slowly varying functions of time, (Xia and Espy-Wilson, 2000). This observation has been employed in setting continuity constraints in order to eliminate wrong formant estimates coming from the short-time segments. The following algorithms imposing formant continuity were found successful:

- Non-linear smoothing: Reliable formant estimates from the neighbor regions are used to approximate the deviating estimate, (Schafer and Rabiner, 1970),

- An extension of reliable formant estimates from strong vocalic areas (anchor frames), (McCandless, 1974),

- Global statistical criteria: HMM is used to find the best overall fit of formant trajectory to the speech signal, (Kopec, 1986). States of the multi-formant model comprise vectors defining possible formant configurations. Continuity constraints are introduced by the transition probability matrix of the HMM. Formant tracking is conducted by the forward-backward algorithm.

- Dynamic programming: Formant candidates are represented by complex roots of the linear predictor. For each frame, local costs of all possible mappings of the candidates to formant frequencies are computed. The costs are determined based on the empirical knowledge of typical intervals of formant occurrences, average locations of formant occurrences, and formant bandwidths (narrow bandwidths are preferred). A modified Viterbi algorithm is used to find the path through the possible mappings yielding minimal costs, while continuity constrains are imposed, (Talkin, 1987), (Xia and Espy-Wilson, 2000).

Various methods were developed for estimating formant bandwidths from the short term spectrum of speech signal, (Dunn, 1961). In the case of LPC-based formant tracking, the bandwidths can be estimated directly from the roots of the all-pole filter. Let p_d be a pole in the z-plane and T be the length of the sampling period:

$$p_d = e^{pT} = e^{(\sigma_p + j\omega_p)T} = e^{\sigma_p T} e^{j\omega_p T}. \tag{4.10}$$

The first term on the right side of (4.10) represents a magnitude of p_d, $R_d = e^{\sigma_p T}$. From the analogy with poles in the *Laplace* plane, (Smith, 2006), (Sovka and Pollák, 2001), the following relation between R_d and the bandwidth B can be found:

$$B = \frac{\ln R_d}{-\pi T}. \tag{4.11}$$

In a recent study, (Deng *et al.*, 2006), the performance of the formant tracker implemented in WaveSurfer, (Talkin, 1987), was compared to the state-of-the-art formant tracker employing a sophisticated model of speech production, (Deng *et al.*, 2004), on the formant-labeled reference database. Both algorithms showed similar performance for vowels, semivowels, and nasals. Due to its availability and good performance on sonorant speech, WaveSurfer was chosen as a tool for formant tracking experiments presented in the remainder of the thesis. The algorithm employs linear prediction analysis[5] and dynamic programming, applied in the manner as discussed earlier in this section.

4.2.1 Error Ellipses

It is common to display mean vowel locations in the F_1–F_2 plane, see Fig. 3.3. To depict the spread of observations contributing to the estimated mean vowel location, error ellipses can be used, (Čmejla and Sovka, 2002). Error ellipses represent an analogy of intervals bounded by the standard deviation σ in 1–D standard distributions. While in 1–D standard distribution the interval $(\mu - \sigma; \mu + \sigma)$ covers 68.2 % of samples centered around the distribution mean μ, in the 2–D plane, the error ellipse covers 39.4 % of samples. Orientation and length of the error ellipse axes can be found as follows. Let $\boldsymbol{C}_{xy}$ be a covariance matrix:

$$\boldsymbol{C}_{xy} = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{yx} & \sigma_y^2 \end{bmatrix}, \tag{4.12}$$

where

$$\sigma_x^2 = \frac{1}{N}\sum_{k=1}^{N}\left(x_k - \bar{X}\right)^2,\ \sigma_{xy} = \sigma_{yx} = \frac{1}{N}\sum_{k=1}^{N}\left(x_k - \bar{X}\right)\left(y_k - \bar{Y}\right), \tag{4.13}$$

$\bar{X}$, $\bar{Y}$ are estimated means of the distributions generating samples x_k and y_k, respectively, and N is the number of samples. The square roots of the eigenvalues[6] of $\boldsymbol{C}_{xy}$ are equal to the lengths of the error ellipse axes, and the corresponding eigenvectors determine the error ellipse axis directions, (Jones, 1999).

Let $\boldsymbol{e}$ be an eigenvector and let λ be an eigenvalue. To find the eigenvalues and eigenvectors, the following equation must be solved:

$$\boldsymbol{e} \cdot \boldsymbol{C}_{xy} = \lambda \cdot \boldsymbol{e}. \tag{4.14}$$

Using the identity matrix $\boldsymbol{I}$, the equation can be rewritten:

$$\boldsymbol{e} \cdot \boldsymbol{C}_{xy} = \boldsymbol{e} \cdot (\lambda \cdot \boldsymbol{I}), \tag{4.15}$$

$$\boldsymbol{e} \cdot (\boldsymbol{C}_{xy} - \lambda \cdot \boldsymbol{I}) = 0. \tag{4.16}$$

For non-trivial $\boldsymbol{e}$, the eigenvalues are found by solving:

$$\begin{vmatrix} \sigma_x^2 - \lambda & \sigma_{xy} \\ \sigma_{yx} & \sigma_y^2 - \lambda \end{vmatrix} = 0. \tag{4.17}$$

[5] The default parameter setting of the formant tracker was used in the analyses: 12^{th} order of LPC, window length 50 ms, window step 10 ms.

[6] An eigenvector of a given linear transformation is a vector that is, as a result of applying the transformation, only scaled by a constant (eigenvalue), while its direction remains preserved, Aldrich (2006).

Expanding the determinant yields a 2^{nd} degree polynomial of λ, called the characteristic polynomial of the matrix. Roots of the characteristic polynomial are the eigenvalues λ_1, λ_2. The corresponding eigenvectors $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ can be determined by finding non-trivial solutions of:

$$\begin{bmatrix} \sigma_x^2 - \lambda_n & \sigma_{xy} \\ \sigma_{yx} & \sigma_y^2 - \lambda_n \end{bmatrix} \cdot \begin{bmatrix} e_{n1} \\ e_{n2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{4.18}$$

The eigenvectors specify the direction of the ellipse axes. If the lengths of the axes are set to $\sqrt{\lambda_1}$ and $\sqrt{\lambda_2}$, the ellipse is called 1–σ or *standard* ellipse, covering 39.4 % of the observed samples. To reach different coverages, the axes lengths are multiplied by an appropriate factor (e.g. (factor/coverage): 2/86.5 % (2–σ ellipse), 2.447/95.0 %, 3/98.9 % (3–σ ellipse), (Mehaffey, 2007)).

4.3 Vocal Intensity

Considerable differences in vocal intensity can be observed when comparing neutral and Lombard speech, see Sec. 3.2.1. In all speech databases used in this thesis, the gain of the microphone preamplifier was adjusted throughout the recording sessions to exploit the dynamic range of the analog-to-digital converter on one side, and to avoid signal clipping on the other side. Therefore, it is impossible to estimate vocal intensity directly from the amplitude of the recorded signal. However, in the case that the ambient noise occurring during the database recording could be considered stationary, vocal intensity changes are proportional directly to the changes of utterance SNRs.

4.3.1 *SNR* Estimation

SNR was estimated using an arithmetical segmental approach, (Pollák, 2002), and voice activity detector (VAD) based on differential cepstral analysis, (Vondrášek and Pollák, 2005). For each short-time frame, the power σ_i^2 is calculated by summing squared frame samples and dividing them by the frame length. VAD is used to estimate the location of frames containing speech. If the i-th segment contains speech, $VAD_i = 1$, otherwise $VAD_i = 0$. In the non-speech segments, the noise power $\hat{\sigma}_{n,i}^2$ is estimated from σ_i^2 and the previous estimates of noise using exponential approximation. In the segments containing speech, the noise power estimate is kept equal to the estimate from the last non-speech segment:

$$\begin{aligned} \hat{\sigma}_{n,i}^2 &= p \cdot \hat{\sigma}_{n,i-1}^2 + (1-p) \cdot \sigma_i^2, \quad VAD_i = 0, \\ \hat{\sigma}_{n,i}^2 &= \hat{\sigma}_{n,i-1}^2, \qquad\qquad\qquad\quad VAD_i = 1. \end{aligned} \tag{4.19}$$

Assuming that the speech and noise components are uncorrelated, the power of speech component $\hat{\sigma}_{s,i}^2$ is estimated by subtracting the noise power from the power of the mixture of speech and noise σ_i^2:

$$\hat{\sigma}_{s,i}^2 = \sigma_i^2 - \hat{\sigma}_{n,i}^2, \quad VAD_i = 1. \tag{4.20}$$

The arithmetical segmental SNR is then estimated for segments j containing speech:

$$SNR = 10 \log \sum_{j=1}^{N} \frac{\hat{\sigma}_{s,j}^2}{\hat{\sigma}_{n,j}^2} \text{ (dB)}, \tag{4.21}$$

where N is the number of frames containing speech. For *SNR* extraction, the SNR Tool, (Vondrášek, 2007), implementing the segmental approach and differential cepstral VAD was used.

4.4 Spectral Slope

The spectral slope of the short-time speech spectra varies significantly with talking styles. As already discussed in Sec. 3.2.4, the slope variations are caused predominantly by changes in the shape of glottal pulses. For the estimation of the glottal waveforms, inverse filtering based on linear prediction is typically used[7]. In the first step, vocal tract parameters are estimated in the periods of glottal closure, since then the speech waveform can be considered a freely decaying oscillation affected only by the vocal tract and radiation by lips, (Cummings and Clements, 1990). Subsequently, the glottal waveform is estimated by filtering the speech signal by the inverse of vocal tract transfer function.

In (Childers and Lee, 1991), a two-pass method for glottal inverse filtering was proposed, see Fig. 4.12. When comparing error function of the linear prediction (LP) and the electroglottographic signal[8] (EGG), it was found that peaks in the error function occur nearly simultaneously with the negative peaks of the differentiated EGG. In the first pass of the method, a fixed-frame LP analysis is performed. In the voiced speech segments, the LP error function comprises peaks occurring in the intervals of the pitch period. The peaks are located and used as indicators of glottal closure. In the second pass, a pseudo-closed interval starting right after the instant of the peak and lasting approximately 35 % of the pitch period is used for a pitch-synchronous LP analysis to get a more accurate estimate of the vocal tract transfer function. Formant structure is then estimated from the LP roots by applying a set of empirical rules. The refined vocal tract transfer function is used for the second-pass glottal inverse filtering. Eventually, the waveform is integrated to remove the effect of the lip radiation.

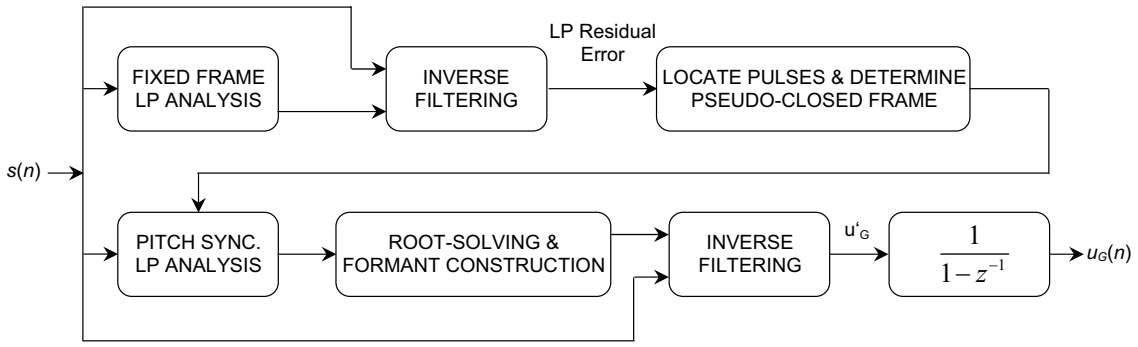

Figure 4.12: Two-pass method for glottal inverse filtering, after (Childers and Lee, 1991).

To estimate the spectral slope of the glottal pulses, it is not necessary to extract the glottal pulse waveform. As mentioned in the beginning of this section, variations of the spectral slope in the short-time speech spectra are caused almost exclusively by the changes in the glottal waveform. Considering the model of speech production discussed in Sec. 3.1, the spectral slope of the glottal waveform and the spectral slope of the speech spectrum differ by a constant +6 dB/oct introduced by the lip radiation.

For the spectral slope analyses conducted in this thesis, the short time spectrum (logarithmic frequency and amplitude axes) was approximated by a straight line, following (Stanton *et al.*, 1988) and (Summers *et al.*, 1988). The linear function modeling the spectral slope was obtained from linear regression, Stephens (1998). In the two-parameter regression analysis, the relation between the

[7]Direct extraction of the glottal waveform is also possible. In (Monsen and Engebretson, 1977), speakers spoke into the reflectionless metal tube, which acted as a pseudoinfinite termination of the vocal tract. The reflectionless termination significantly reduced vocal tract resonances, hence, the signal sensed at the end of the tube corresponded to the vocal tract excitation.

[8]Electroglottographic signal is obtained by measuring the electrical impedance across the throat.

variables x and y is modeled:

$$y_i = \alpha + \beta x_i + \varepsilon_i, \tag{4.22}$$

where ε_i is a noise term with zero mean. Parameters of the line are estimated by means of least squares:

$$\hat{\beta} = \frac{N \sum_{i=1}^{N} x_i y_i - \sum_{j=1}^{N} x_j \sum_{k=1}^{N} y_k}{N \sum_{l=1}^{N} x_l^2 - \sum_{m=1}^{N} x_m \sum_{n=1}^{N} x_n}, \tag{4.23}$$

$$\hat{\alpha} = \frac{\sum_{i=1}^{N} y_i - \hat{\beta} \sum_{j=1}^{N} x_j}{N}, \tag{4.24}$$

where N is the number of samples. Because the regression was performed in the log–log domain, $x_i = log(f_i)$, $y_i = log(H_i)$.

4.5 Duration

Phoneme and word durations in neutral and Lombard speech may differ, see Sec. 3.2.5. In the analyses conducted in this thesis, the duration differences were evaluated as

$$\Delta = \frac{T_{LE} - T_{neutral}}{T_{neutral}} \cdot 100 \ (\%), \tag{4.25}$$

where T_S represents the average phoneme or word duration in the scenario S.

4.6 Feature Analyses Metrics

This section presents metrics used for evaluating results of feature analyses.

4.6.1 Weighted Means and Deviations

In some of the analyses, parameter estimates were extracted from segments of varying lengths (e.g., spectral slopes or formant locations and bandwidths). It seems natural that the parameter values extracted from longer time segments should contribute more to the parameter mean and deviation estimates than those coming from shorter segments. Hence, in the analyses employing varying segments, weighted means and deviations were calculated:

$$\overline{X}_w = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{j=1}^{N} w_j}, \tag{4.26}$$

$$\hat{\sigma}_w = \sqrt{\frac{\sum_{i=1}^{N} w_i \left(x_i - \overline{X}_w\right)^2}{\sum_{j=1}^{N} w_j}}, \tag{4.27}$$

where w_i is the weight (typically a time duration) of the i-th sample x_i and N is the number of samples.

4.6.2 Student's t-Test

To evaluate whether two sets of observations come from the same distribution, i.e., whether the sample means are equal from the statistical point of view (although may be numerically different), the independent Student's t-test was used. Since the means to be compared were usually calculated from different number of samples, the independent test for unequal sample sizes, Stephens (1998), was carried out. The t value was determined:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\hat{\sigma}_{\overline{X}_1 - \overline{X}_2}}, \tag{4.28}$$

where $\overline{X}_1$ and $\overline{X}_2$ are the compared mean estimates, and $\hat{\sigma}^2_{\overline{X}_1 - \overline{X}_2}$ is the unbiased estimator of variance

$$\hat{\sigma}_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{(N_1 - 1)\,\hat{\sigma}_1^2 + (N_2 - 1)\,\hat{\sigma}_2^2}{N_1 + N_2 - 2}\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}, \tag{4.29}$$

$N_{1,2}$ are the numbers of samples in the measurement sets, $N_1 - 1$ and $N_2 - 1$ are degrees of freedom for the sets, and $N_1 + N_2 - 1$ is the total number of degrees of freedom. The null hypothesis that the two means do not differ was tested at the significance level $\alpha = 0.05$. If the t value exceeded the level corresponding to α, the null hypothesis was rejected and the means were considered statistically different.

4.6.3 Confidence Intervals

To allow for easy comparison of more than two estimates of a given parameter, 95% confidence intervals were determined for each estimation. The 95% confidence interval was constructed, Stephens (1998):

$$P\left(\bar{X} - 1.96\frac{\hat{\sigma}}{\sqrt{N}} \leqslant \mu \leqslant \bar{X} + 1.96\frac{\hat{\sigma}}{\sqrt{N}}\right) \approx 0.95, \tag{4.30}$$

where $\bar{X}$ is the sample mean, $\hat{\sigma}$ is the sample standard deviation, N is the number of samples, and μ is the unknown real population mean. The expression (4.30) says that the real mean μ can be expected to lie in the given confidence interval with confidence level of 95 %.

In the case of presenting results together with confidence intervals, two estimated means were considered statistically different if each of them lied outside the confidence interval of the other mean.

4.7 Recognition Setup

The structure and algorithms employed in HMM-based recognition systems were discussed in detail in Sec. 2. This section briefly summarizes the setup of the digit and LVCSR recognizers employed in the experiments.

4.7.1 Digit Recognizer

The digit recognizer was implemented in HTK (Hidden Markov Model Toolkit), Young *et al.* (2000). The acoustic models had the following architecture:

- 43 context-independent left to right monophone models and two silence models (short pause and long pause), (Novotný, 2002),
- Each monophone model and the long pause model comprised 3 emitting GMM states, the short pause model contained one emitting GMM state,

- Each GMM was represented by a 39–D multiple-mixture Gaussian distribution function. The dimensions modeled 13 static, 13 dynamic, and 13 acceleration coefficients.
- Each GMM dimension employed 32 Gaussian mixtures.

In the initial experiments, MFCCs were used for the feature extraction, see Sec. 2.5. First, an energy coefficient and cepstral coefficients C_1–C_{13} formed the static feature vector. In the later experiments, C_0–C_{13} (C_0 representing the frame energy) were used. Both configurations displayed comparable performance. The cepstral coefficients were extracted from the bank of 26 triangular filters. Window length of 25 ms and 10 ms overlap were used for the signal segmentation in the majority of experiments. Pre-emphasis $\alpha = 0.97$, see Eq. (2.33), was applied to the segments before calculating the short-time spectrum and extracting the cepstral coefficients.

The monophone acoustic models were trained in four steps. First, global means and variances of the training speech data were calculated for each feature from the 39-D feature vector and used as the initial parameters of the monophone and silence models, i.e., each of the monophone and silence models was initialized by the same global values. This approach is called *flat start*. In this stage, each GMM dimension comprised only a single Gaussian component. Second, Baum–Welch expectation-maximization algorithm, Young *et al.* (2000), was applied to re-estimate the model parameters, given the training speech files and corresponding phonetic transcriptions. The phonetic transcriptions can be obtained from manual annotations of the speech recordings, if available, or from orthographic transcriptions of the speech recordings by picking one of the pronunciation variants provided by the lexicon. In this work, the latter approach was used, picking the first pronunciation variant of each word. The Baum–Welch re-estimation was repeated several times.

In the third stage, the actual monophone models were used to *realign* the training data and create new transcriptions. Given the orthographic transcription for each speech file, maximum likelihood pronunciation variants were searched for the transcription words using the Viterbi algorithm. If the pronunciation variants did not yield a likelihood exceeding a given threshold (pruning threshold), the speech file was excluded from the training set. Otherwise, the maximum likelihood pronunciation variants replaced the former phonetic transcription of the speech file. After the realignment, model re-estimation was conducted several times.

Finally, the single Gaussian components were split into Gaussian mixtures. Two alternatives of mixture-splitting, *one-step* and *progressive mixture splitting* were employed in the training process[9]. In the *one-step* approach, the former Gaussian component was split into the desired number of mixtures in one step, followed by multiple model re-estimations. In *progressive mixture splitting*, the Gaussian was split into two Gaussian mixtures, followed be several re-estimations. Subsequently, the mixture doubling and model re-estimation was repeated until the desired number of mixtures (power of two) was reached. The performance of *one-step* and *progressive mixture splitting* is compared in Sec. 9.4.1.

In the experiments, the training data comprised either gender-dependent or gender-independent utterances, yielding gender-dependent or gender-independent acoustic models, respectively. For the decoding, the Viterbi algorithm was used. In the digit recognition experiments, the language model comprised 10 Czech digits in 16 pronunciation variants with uniform probabilities of transitions between the words.

4.7.2 LVCSR Recognizer

The LVCSR system[10], (Nouza *et al.*, 2005), employed a multiple-pronunciation vocabulary containing over 300.000 words and a bigram language model trained on the 4 GB newspaper corpus. The recognizer's front end employed MFCC features and the segmentation 25/10 ms. The acoustic

[9]Performance of the mixture splitting approaches was compared in (Bořil and Fousek, 2007).

[10]The LVCSR system was kindly provided for the experiments by Prof. Jan Nouza, Technical University of Liberec.

model comprised 48 monophone, pause, and noise models. Structure of the models was similar to the digit recognizer. Sets of both gender-independent and gender-dependent models were available for the experiments.

From the beginning of the training, the model states comprised the desired number of mixtures (up to 100), which were no more split. For the training, the models were initialized by Viterbi training, Young *et al.* (2000). Besides the speech signals and orthographic transcriptions, the training data comprised phonetic transcriptions and corresponding time labels of the phone occurrences. To obtain parameter estimates for each of the model states, the segments associated to the given model were divided into successive uniform portions assigned to the successive model states. The state means and variances were estimated by averaging the corresponding portions. Subsequently, Viterbi algorithm was used for the alignment and re-estimation of the model parameters given the new segment boundaries. The alignment and model update was repeated until the alignment likelihood stopped rising. Once the models were initialized, 15–20 Baum–Welch re-estimations were carried out.

4.7.3 Recognition Evaluation

To assess performance of ASR systems in the presented experiments, the word error rate (WER) was used. WER evaluates the difference between the string of words returned by the recognizer and the correct transcription. First, an optimal mapping between the hypothesized and correct word sequence involving minimum word substitutions (S), insertions (I) and deletions (D) is searched, Jurafsky and Martin (2000). WER is then calculated:

$$WER = \left(\frac{D + S + I}{N} \right) \cdot 100 \ \ (\%)\,, \tag{4.31}$$

where N is the number of words in the correct transcription. To examine the statistical significance of WER differences across experiments, confidence intervals were determined. Let $\{u_i\}$ be a sequence of words to be recognized. Let X_i be a random variable:

$$X_i = \begin{cases} 0, \text{ correct recognition of } u_i, \\ 1, \text{ incorrect recognition of } u_i. \end{cases} \tag{4.32}$$

Let N be a number of words and p be the unknown real error rate. If the recognition errors are independent events, it can be assumed that the sum $S = \sum_{i=1}^{N} X_i$ follows a binomial distribution $B(N, p)$, (Gillick and Cox, 1989). The maximum likelihood estimate of p is

$$\hat{p} = \frac{S}{N} = \frac{\sum_{i=1}^{N} X_i}{N}. \tag{4.33}$$

The unknown variance of p is

$$\sigma^2\,(p) = \frac{p\,(1-p)}{N}. \tag{4.34}$$

The 95% confidence interval can be estimated:

$$P\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}\,(1-\hat{p})}{N}} \leqslant p \leqslant \hat{p} + 1.96\sqrt{\frac{\hat{p}\,(1-\hat{p})}{N}} \right) \approx 0.95. \tag{4.35}$$

Following (Mokbel and Chollet, 1995), two results are considered significantly different if each one of them is outside the confidence interval of the other result.

Chapter 5

Design of Czech Lombard Speech Database (CLSD'05)

Speech databases acquired in real conditions provide a valuable resource for the design and testing of ASR systems. On the other hand, in the case of louder backgrounds (crowded places, moving car, airplane cockpits) it may be problematic to reliably analyze speech production variations induced by noise and their impact on ASR performance separately from the impact of the signal contamination by noise. Also, assuring similar recording conditions across speakers and appropriate speaker's reactions to the actual noise may be an issue in the real conditions.

In this chapter, the design of Czech Lombard Speech Database (CLSD'05)[1] comprising speech uttered in neutral and simulated noisy conditions is presented. The main goal of the database is to provide:

- Neutral and LE speech recordings with high SNR,
- Defined spectral properties and levels of the noisy backgrounds,
- Communication between speaker and operator in noise, motivating speakers to react appropriately to the actual noise,
- Parallel neutral and LE utterances from the same speakers for analyses of feature variations on the speaker/utterance level,
- Extensive number of utterances containing digits, commands, and complex sentences to allow for statistically significant small and LVCSR recognition experiments.

In databases of simulated LE, speakers are usually provided noisy background through headphones while their speech is sensed by close talk microphone, yielding high SNR of the recordings, e.g., (Junqua and Anglade, 1990), (Chi and Oh, 1996), (Hansen and Bou-Ghazale, 1997). To introduce the communication factor motivating speakers to produce intelligible speech in noise, this 'standard' setup was extended in the presented work for an operator listening to the utterances while being exposed to the same noise. The utterance could not be understood or was misheard by the operator, the speaker was asked to repeat it[2].

[1] The CLSD'05 design and acquisition were published in (Bořil and Pollák, 2005b).

[2] A similar approach can be found in (Korn, 1954) and (Webster and Klumpp, 1962). In the latter study, ambient noise was produced from the speaker array instead of headphones. It is noted that both studies focused exclusively on finding the Lombard function (i.e., the dependency between noise intensity and corresponding vocal effort in speakers) and did not conduct an analysis of other speech features. At the time of the submission of this thesis, a speech acquisition employing a pair speaker–listener exposed to the same noise background was adopted for the purposes of LE speech

For the purpose of evaluation of speech feature variations on the speaker level, each speaker was recorded both in neutral and simulated noisy conditions. Speakers produced the same utterances in both scenarios (parallel utterances) to allow for context dependent feature analyses.

To obtain statistically significant results from the analyses and recognition experiments, an extensive number of small vocabulary items (digits, orders) as well as large vocabulary items (phonetically rich, often complex, sentences) were included in the corpus.

5.1 Recording Setup

During the recording, a speaker was asked to read utterances from the LCD panel. Speech was recorded via close talk and hands free (middle talk) microphones. In the simulated noisy conditions, speakers wore headphones. Closed headphones were used to eliminate crosstalk of the noise from headphones to the close talk microphone. Speech obtained by the close talk microphone was fed back to the speaker (speech feedback) to suppress the sound attenuation caused by wearing headphones. In the beginning of the noisy recording (Lombard scenario), the speaker was asked to adjust the speech feedback to a comfortable level (i.e., to the level where they would hear themselves as if they were not wearing the headphones[3]). During the recording, noise samples were mixed with the speech feedback and reproduced to the speaker by the headphones. Simultaneously, an operator was qualifying intelligibility of the speaker's utterances while being exposed to the same noise as the speaker. The utterances mixed with noise were delivered to the operator through headphones. The monitored speech in the operator's headphones was attenuated to simulate fading of speech intensity due to the distance between the speaker and the operator. If the speech in noise was not understood well, the operator asked for it to be repeated. The recording setup is outlined in Fig. 5.1.

The recording set consisted of 2 closed headphones AKG K44, close talk microphone Sennheiser ME-104, and hands-free microphone Nokia NB2. These microphones were similar to those used for the acquisition of Czech SPEECON, (Iskra *et al.*, 2002).

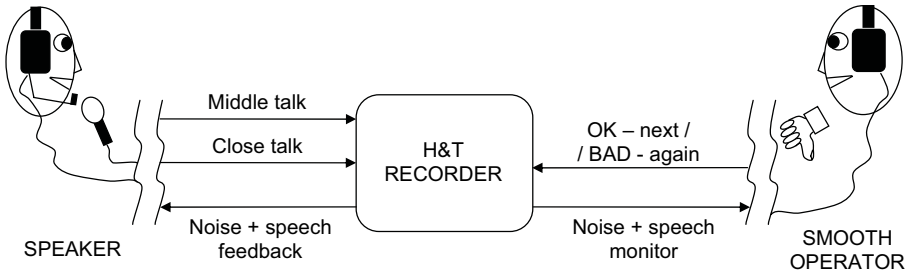

Figure 5.1: Recording setup.

5.2 SPL Adjustment

In the beginning of each Lombard recording, it was necessary to adjust the level of the reproduced background noise. For this purpose, a transfer function between the effective voltage of the sound card open circuit V_{RMS_OL} and SPL in the headphones was measured on a dummy head, see Fig.

analysis in (Patel and Schell, 2008). No study employing this concept for the analysis of the impact of LE on ASR is known to the author of this thesis.

[3]Similar scheme can be found in (Bond *et al.*, 1989). In (Steeneken and Verhave, 2004), a speech feedback eliminating the frequency-dependent attenuation caused by wearing headphones was designed. It is noted that in this case, the speech feedback was a component of active noise reduction system and was not used for acquisition of LE speech.

5.2. The required noise level could be adjusted by setting the corresponding V_{RMS_OL}. A noise level of 90 dB SPL and a virtual distance of 1–3 meters were used for the Lombard speech recordings. In some cases, the settings had to be modified according to the particular speaker's capabilities. The noise reproduction was interrupted between the consecutive items to be read. The recording session usually did not exceed 20–30 minutes per scenario, with refreshment pauses included.

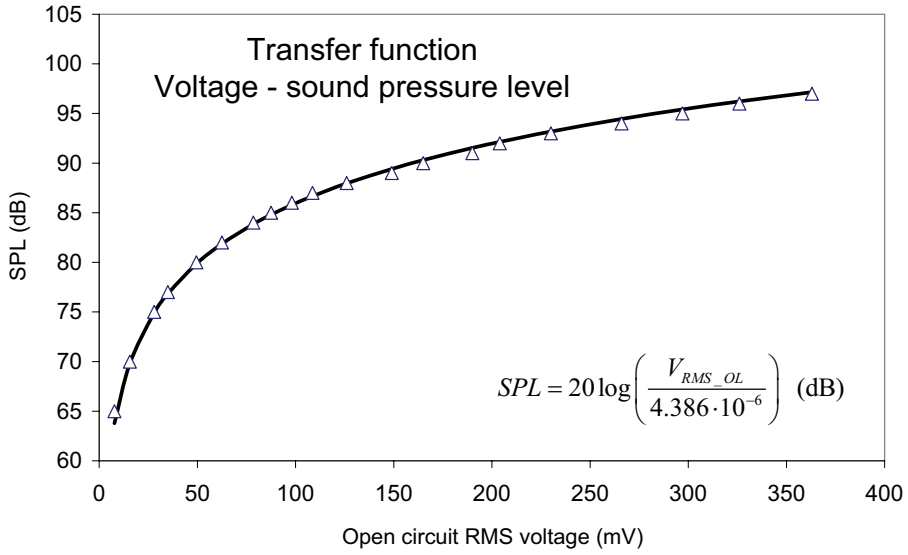

Figure 5.2: V_{RMS_OL}–SPL dependency.

5.3 Noise Samples

Noises recorded in the car cabin and artificial band-noises interfering with typical locations of speech fundamental frequency and first formants were employed in the Lombard recordings. A total 25 quasi-stationary car noises selected from the CAR2E database, (Pollák *et al.*, 1999), and 4 band noises (62–125, 75–300, 220–1120, 840–2500 Hz) were used for the noisy backgrounds. The car noise samples were about 14 sec long each, the band-noises were 5 sec long. The 29 noises were assigned to the session prompts periodically, one noise sample per prompt. For the cases where the utterance would have exceeded the noise sample length, the noise sample was looped. All noise samples were RMS normalized.

5.4 Recording Studio

For the acquisition of the CLSD'05 database, the H&T Recorder was developed[4]. The H&T Recorder supports two-channel recording and separate noise/speech monitoring for the speaker and the operator.

In the main application window, see Fig. 5.3, a list of session prompts is displayed. Besides consecutive passing through the prompt list, the operator can jump to any prompt of choice or use the 'first/previous/next/last unrecorded' selector (the bottom row of buttons in Fig. 5.3). For each item of the prompt list, a separate noise sample can be assigned from the noise list. Levels and waveforms

[4]The H&T Recorder was implemented in .NET by Tomáš Bořil, Czech Technical University in Prague.

of the recorded signals from both microphone channels are displayed in the application window so any unintended signal overdriving or utterance cutting can be easily revealed.

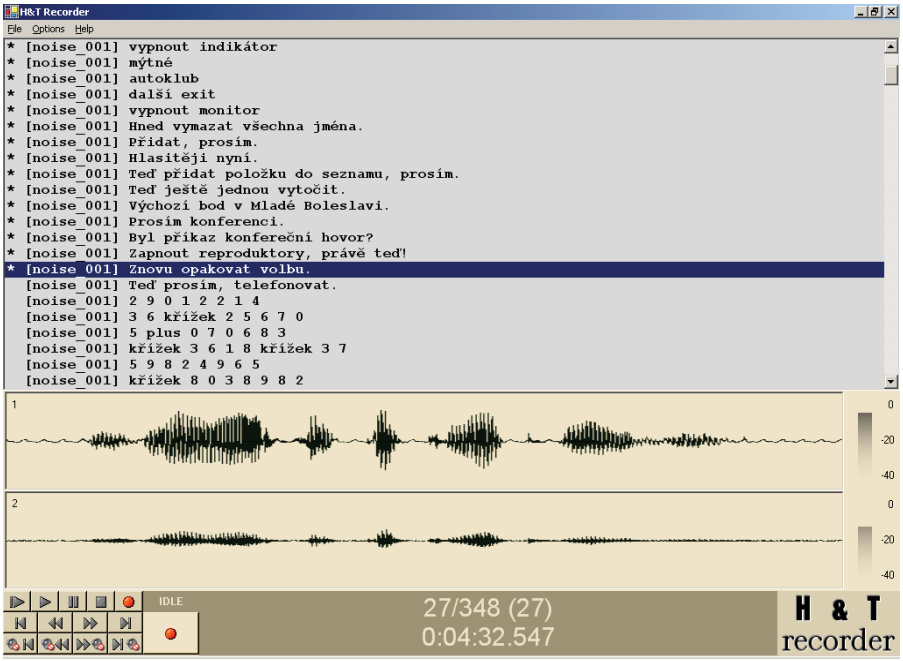

Figure 5.3: H&T Recorder – application window.

In the 'Options' menu, session sample frequency, number of channels (1/2), level of speech and noise signals sent to speaker's headphones, and speaker–operator distance can be chosen. Every utterance is saved to the hard disc immediately after being recorded. For each recorded utterance, a label file is created. The format of the label file originates from the SAM label file used in SPEECON, (Iskra *et al.*, 2002), comprising information about the actual speaker, recording scenario, orthographic and phonetic transcription, etc. For the purposes of CLSD'05, the SAM label file was extended for items describing the actual simulated noisy conditions, see Table 5.1.

SAM label	Format	Description	Notice
NTY	%s	Noise type	Filenames including noise description code
NLV	%f	Noise level	Level of noise in headphones (dB)
DIS	%f	Distance	Speaker – operator distance (meters) ⇒ speech attenuation in operator's monitor

Table 5.1: SAM label file extension in CLSD'05.

5.5 Corpus

CLSD'05 comprises utterances from 26 speakers (12 female, 14 male) participating both in neutral and noisy scenarios. Recording sessions typically contain 205 utterances (in average 780 words) per speaker and scenario, which represents about 10–12 minutes of continuous speech. The utterance files are stored in a raw file sampled by 16 kHz, in a 16-bit sample format.

In order to represent the whole phoneme material occurring in the Czech language and enable LVCSR experiments, 30 phonetically rich sentences (often complex) appear in each session. To allow for statistically significant small vocabulary recognition experiments, 470 repeated and isolated digits were included in each session. In the case of SPEECON, 40 digits are available per session. A typical content of the CLSD'05 session is shown in Table 5.2.

Corpus contents	Corpus/item id.	Number
Phonetically rich sentences	S01 – 30	30
Phonetically rich words	W01 – 05	5
Isolated digits	CI1 – I4, 30 – 69	44
Isolated digit sequences (8 digits)	CB1 – B2, 00 – 29	32
Connected digit sequences (5 digits)	CC1 – 4, C70 – 99	34
Natural numbers	CN1 – N3	3
Money amount	CM1	1
Time phrases; T1 – analogue, T2 – digital	CT1 – T2	2
Dates: D1 – analogue, D2 – relative and general date, D3 – digital	CD1 – D3	3
Proper name	CP1	1
City or street names	CO1 – O2	2
Questions	CQ1 – Q2	2
Special keyboard characters	CK1 – K2	2
Core word synonyms	Y01 – 95	89
Basic IVR commands	101 – 85	
Directory navigation	201 – 40	
Editing	301 – 22	
Output control	401 – 57	
Messaging & Internet browsing	501 – 70	
Organizer functions	601 – 33	
Routing	701 – 39	
Automotive	801 – 12	
Audio & Video	901 – 95	

Table 5.2: Typical CLSD'05 session content.

5.6 Attenuation by Headphones

Altering the level of speech feedback (sidetone) has a similar effect on the speaker's vocal effort as altering the level of background noise, (Lane *et al.*, 1961). Hence, the individual adjustments of the speech feedback level in the beginning of each Lombard session might have affected the way speakers reacted to noise. Since the level of the monitored speech in the operator's headphones was derived from the level of the speaker's speech feedback, both speaker and operator were influenced by the initial feedback adjustment. To eliminate the drawback of using the subjectively adjusted feedback, a new method of precise speech feedback adjustment is proposed for the purposes of further recordings[5].

Sound waves propagate to human senses through air vibration and head vibrations (skull bone conduction). If speakers judge dynamics of their own speech, proprioception also contributes to the sound perception, (Lane *et al.*, 1970). It can be expected that wearing closed headphones causes attenuation of the sound passing through the ears while the bone conduction remains intact (although the mass of the head is slightly increased by the headphones)[6]. In several studies, the effect of sound attenuation was reduced by maintaining a gap (about one inch) between the speaker's ears and the

[5]The speech feedback design was published in (Bořil *et al.*, 2006a).

[6]In (Steeneken and Verhave, 2004), a speech feedback was included in the active noise reduction system in order to assure that speakers will hear themselves (without the speech feedback, the system would suppress not only the noisy background but also the speaker's speech). Due to the sound attenuation caused by wearing closed headphones, there is a difference between characteristics of the speech waveform approaching the headphones and characteristics of the

headphones, (Korn, 1954), or by using open-air headphones, (Junqua *et al.*, 1998), (Varadarajan and Hansen, 2006). However, in these setups, the sound crosstalk between the headphones and the close talk microphone can be expected to be considerably higher than in the case of closed headphones. Considering that the noise level used for LE speech recording in the present database was set to 90 dB SPL, the noise crosstalk due to using open-air setup might have significantly contaminated the speech recordings. To prevent this, closed headphones were used for the data acquisition in this thesis. In the remainder of this section, characteristics of the sound attenuation by the AKG K44 headphones are analyzed and a speech feedback transfer function compensating for the attenuation is proposed.

Characteristic of the attenuation by headphones was measured on a dummy head in an anechoic room using Pulse v.8 (Brüel and Kjaer, 2004). The dummy head employed models of auditory canals and complied with the recommendation ITU P.58, (ITU, 1996)[7].

Monaural directional frequency responses were measured on the dummy without and with headphones. Subsequently, the attenuation characteristic was determined as a subtraction of these responses (see Fig. 5.4). The measurement was carried out for angles 0–180°. In the case of 0° the head's nose, and for 90° the head's ear were directed to the source of the measuring noise, respectively. The measurement was not performed for angles greater than 180° as in the anechoic room the sound would propagate to the measured ear only by the dummy head vibrations.

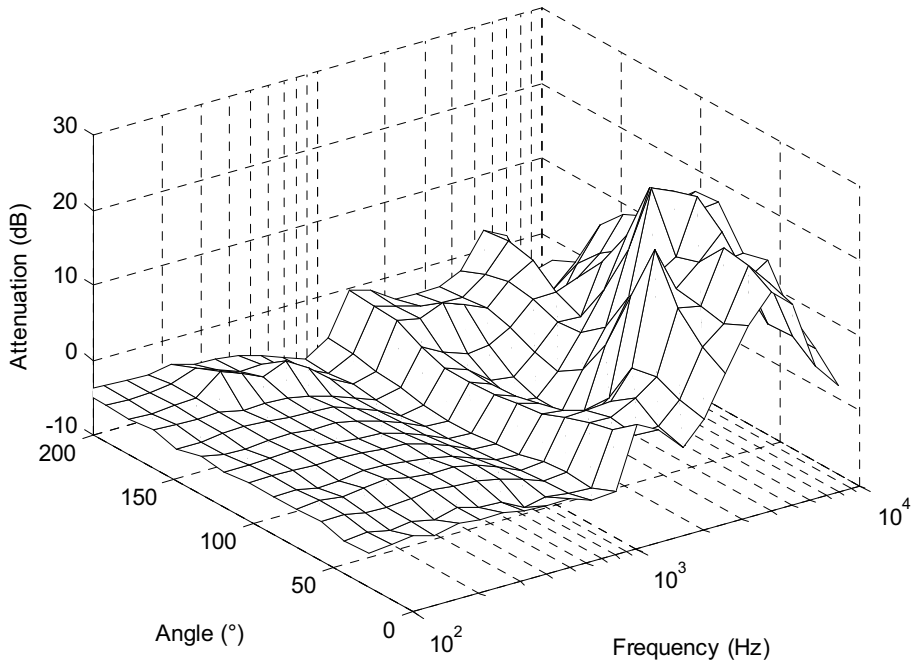

Figure 5.4: Attenuation by headphones – directional characteristics.

Results of the measurement show that the frequency response of the attenuation depends significantly on the direction of the sound source. Directionality of the attenuation is shown in detail for selected frequencies in Fig. 5.5. Since sound waves tend to reflect from barriers, the reflected speech sound will propagate to the ears from various angles and intensities, depending on the room size, sound absorption coefficients of the walls, and speaker's position and orientation in the room. As shown in

waveform sensed inside of the headphones' cavity by the microphone of the noise reduction system. To address this, the authors of the study designed a speech feedback compensating for the attenuation by the headphones' ear muffs.

[7] A perceptual sound attenuation by wearing another type of headphones in a diffuse sound field has been studied in (Arlinger, 1986).

Fig. 5.4 and Fig. 5.5, the attenuation is less significant and also less directional for low frequencies. Actually, some harmonic components are even slightly amplified by the presence of headphones.

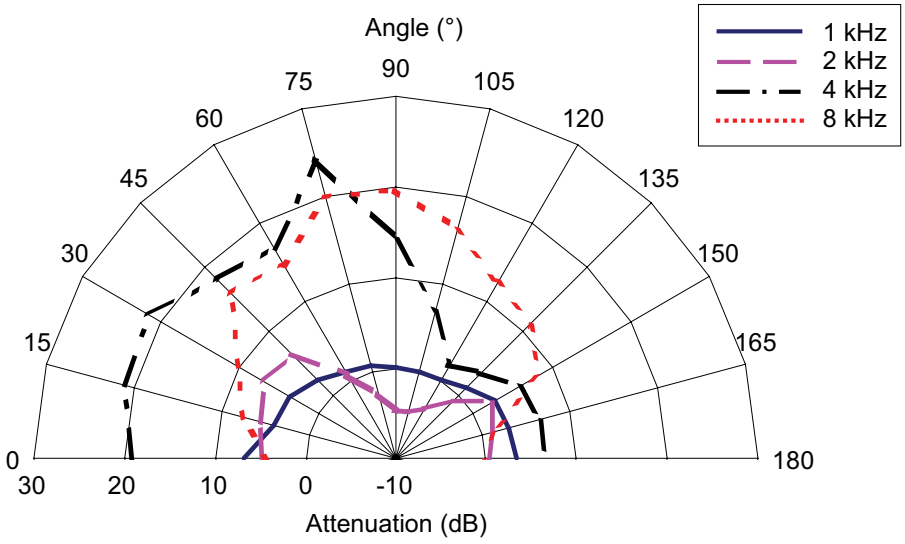

Figure 5.5: Directional characteristics for selected frequencies.

Subsequently, the attenuation characteristic was measured in the CLSD'05 recording room. The dummy head was placed in a position where the speakers were seated while being recorded. A loudspeaker was put in front of the dummy head's mouth to provide a sound excitation of the room. Third octave band noises were used for the room excitation in the interval of 80–8000 Hz. Monaural transfer functions were measured on the dummy head without and with headphones. The attenuation characteristic was determined as their difference. The attenuation characteristic for the CLSD'05 recording room is depicted in Fig. 5.6. Selected anechoic room characteristics are also shown for a comparison – 0°, 90°, 180°.

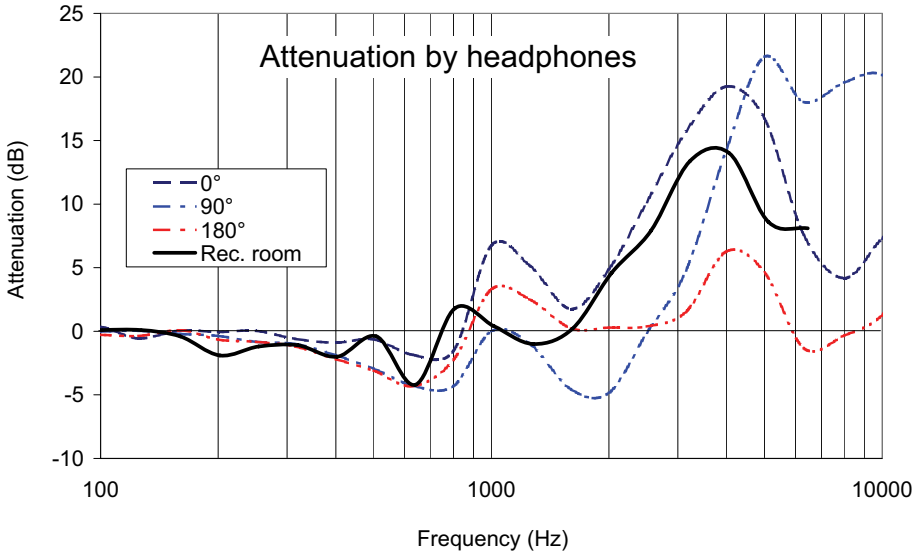

Figure 5.6: Anechoic and CLSD'05 recording rooms – frequency responses of attenuation.

Surprisingly, the attenuation by headphones does not increase monotonously with increasing frequency[8], but rather starts to decrease for frequencies above 4 kHz. From the equal loudness curves, (Fletcher and Munson, 1933), 'Equal-Loudness Contour', it can be seen that the area of maximum sensitivity of human hearing lies in the interval of 3–4 kHz, which is related to the resonance of the auditory canal. Based on the measurement results, it can be assumed that the headphones affect the configuration of the ear resonator, resulting in an upward shift of the resonant frequency. This shift causes a significant attenuation peak in the former resonant area and a significant drop in the area of the new resonance, resulting in a decrease of attenuation for frequencies above the former resonance.

Once the attenuation characteristic for the given room, speaker position, and headphones is found, a compensation function can be determined by inverting the attenuation characteristic. In the case of CLSD'05, the speech feedback compensating the attenuation by headphones would have a transfer function of similar shape as the attenuation curve shown in Fig. 5.6. Since individual sizes and properties of the auditory system differ across speakers, the compensation function derived from the measurements on the dummy head is an approximation of the 'optimal' speaker-dependent attenuation compensation. However, using the general compensation function is a promising way to improve how speakers perceive themselves when wearing closed headphones. Evaluating effectiveness of the proposed compensation feedback in the perceptual tests is a matter of further work not considered in this study.

5.7 Signal Level Reconstruction

During the CLSD'05 recording, it was necessary to modify the gain of the microphone preamplifier from time to time to avoid signal clipping when speakers changed their vocal intensity. As a consequence, it is impossible to determine the vocal intensity directly from the amplitude of the recorded speech signal. In (Pisoni *et al.*, 1985), a similar problem was addressed by recording the speech signal using two channels. In the first channel, the gain was adjusted during the recording to cover the entire dynamic range of the analog-to-digital converter, while in the second channel the gain was kept constant throughout the whole recording to preserve the relative amplitudes between tokens. In CLSD'05, the ambient noise occurring during recording can be considered stationary and, hence, the changes in vocal intensity are proportional to SNRs of the recordings.

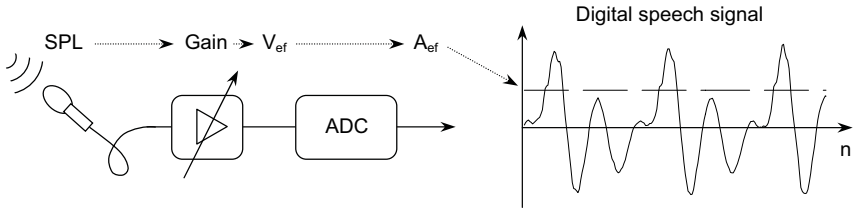

Figure 5.7: Relation of signal level and SPL.

For the purposes of further recordings, an improved one-channel method preserving information about the absolute signal amplitudes is proposed. In the first step, the transfer function V_{ef}/SPL of the close talk microphone is measured for the known microphone preamplifier gain. Secondly, during the speech recording, the calibrated microphone is placed at a selected distance from the speaker's lips. The actual preamplifier gain is stored for each utterance. Knowledge of the actual gain is sufficient for the reconstruction of SPL of the acoustic signal, see Fig. 5.7. Moreover, based on the formula

[8]Intuitively, it could be expected that the higher is the sound frequency, the higher would be the attenuation of the sound passing through a barrier.

derived in (Titze and Sundberg, 1992), the power radiated from mouth can be calculated given the microphone distance from lips and sound SPL measured by the microphone.

Chapter 6

Baseline Experiments on Selected Czech Corpora

In this chapter, three databases of Czech speech recordings are analyzed[1]. Two extensive databases, Czech SPEECON, (Iskra *et al.*, 2002), and CZKCC, (CZKCC, 2004), and the newly established CLSD'05 database were selected for the experiments. CLSD'05 is assumed to capture neutral and Lombard speech due to the definition of its recording setup. Czech SPEECON and CZKCC comprise recordings from quiet and real adverse conditions, therefore they are promising candidates to provide neutral and actual Lombard speech.

6.1 Databases

6.1.1 Czech SPEECON

Czech SPEECON database provides the following content:

- Speakers: 590 adults, 50 children, approximately 30 minutes of speech per speaker,
- Speech signals: 4 channels, $F_s = 16$ kHz, linear PCM coding,
- Content: Phonetically rich items, application-oriented utterances, digit sequences,
- Environments: car, office, public places (living rooms, exhibition areas),
- Recording channels – office and public places: close talk, hands free, directional microphone (placed 1 meter from a speaker), omni-directional microphone (placed 2–3 meters from a speaker),
- Recording channels – car: close talk, hands free, microphone in the closest front corner of the car cabin, distant front corner microphone.

Office and *car* recordings were selected for the baseline experiments. *Office* scenario utterances were recorded in a calm environment and can be assumed to capture neutral speech. *Car* scenario recordings comprise utterances produced in noisy conditions. Different speakers participated in the neutral and noisy conditions.

[1]Results of the analyses were published in (Bořil and Pollák, 2005a).

6.1.2 CZKCC

CZKCC comprises recordings from the car cabin:

- Speakers: 1000 adults,
- Speech signals: 2 channels, $F_s = 48$ kHz, linear PCM coding,
- Content: Phonetically balanced sentences, orders, digit sequences,
- Environments: Standing car/engine off, standing car/engine on, moving car/engine on,
- Recording channels – setup I: Close talk, distant front microphone (above the middle rear view mirror),
- Recording channels – setup II: Two distant front microphones (above the middle rear view mirror).

Standing car/engine off and *moving car/engine on* recordings were selected to represent neutral and noisy speech, respectively. Recordings obtained using the setup I were employed in the experiments. The same speakers participated in the neutral and noisy conditions.

6.1.3 CLSD'05

Structure of CLSD'05 is described in detail in Chap. 5. The database parameters are as follows:

- Speakers: 26 adult speakers,
- Speech signals: 2 channels, $F_s = 16$ kHz, linear PCM coding,
- Content: Phonetically rich items, application-oriented utterances, digit sequences,
- Environments: Office – neutral, simulated noisy conditions,
- Recording channels: Close talk, hands free.

The same speakers participated in the neutral and noisy conditions.

6.2 Feature Analyses

Data from the close talk channel were used in the experiments as they provide signals with the highest *SNR*. In the case of CZKCC, data was down-sampled from 48 kHz to 16 kHz using the SoX tool, (Norskog, 2007). The number of speaker sessions employed in the experiments are shown in Table 6.7, *'Spkrs'*. Microphone channel *SNRs*, fundamental frequency distributions, first four formant positions and bandwidths, and phoneme and word durations were evaluated. For the analyses, framework described in Chap. 4 was used. In the *SNR* and F_0 analyses, all session utterances were processed. In the formant analyses, only digit utterances were employed.

6.3 *SNR*

SNR distributions for the first two microphone channels in neutral and noisy scenarios are shown in Fig. 6.1. In the case of CLSD'05, *N* denotes neutral speech and *LE* stands for Lombard speech. The distribution of means and standard deviations are shown in Table 6.1. In the case of Czech SPEECON and CZKCC, the level of background noise present in the speech signal varied considerably

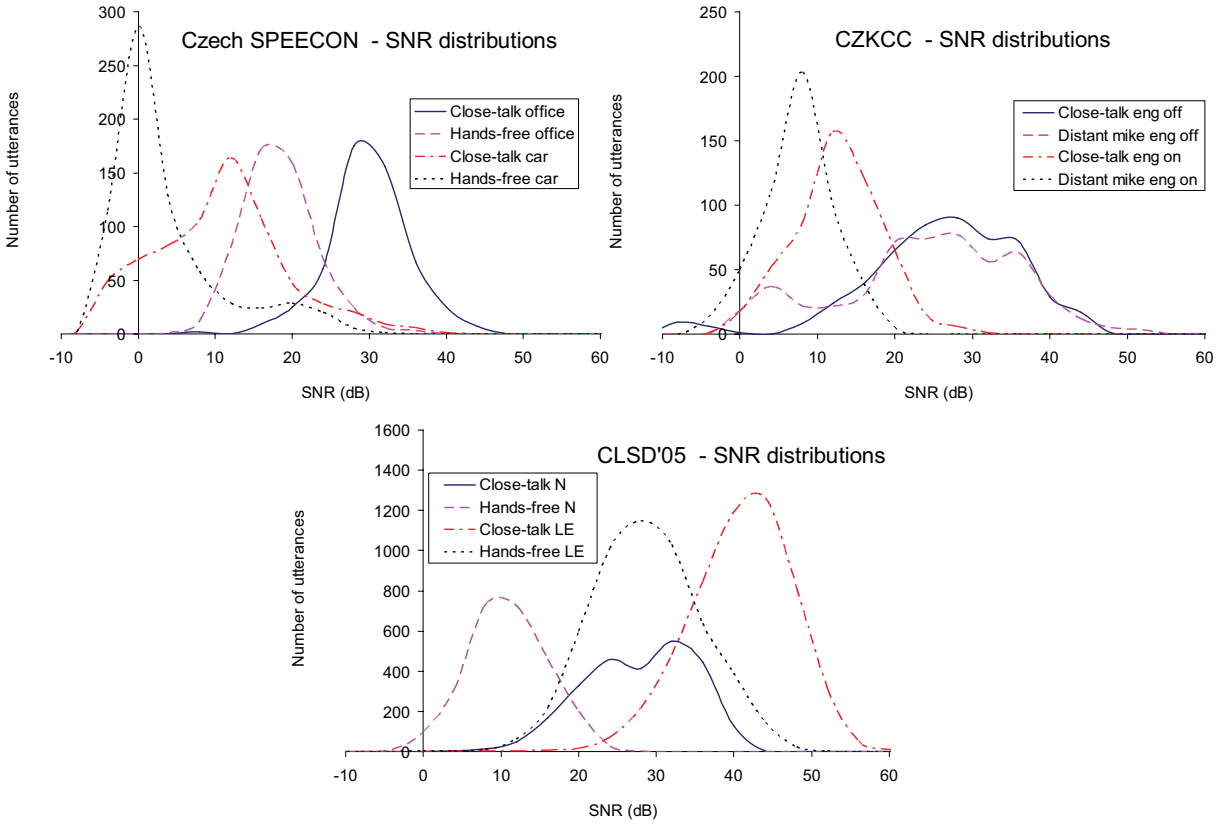

Figure 6.1: SNR *channel distributions: Czech SPEECON, CZKCC, and CLSD'05.*

when comparing *office* and *car*, or *standing car/engine off* and *moving car/engine on* scenarios. In the case of CLSD'05, the ambient noise as sensed by the microphone during the recording of neutral and simulated Lombard speech could be assumed stationary due to the use of closed headphones. Here, the *SNR* histograms can be assumed to correlate with distributions of vocal intensity in neutral and Lombard utterances. During the CLSD'05 acquisition, the level of background noise in the room ranged approximately from 50 to 60 dB SPL. In simulated LE conditions, speakers were exposed to a noise level of 90 dB SPL. As shown in Table 6.1, the average vocal level in speakers increased by approximately 13 dB when switching from neutral to noisy conditions (i.e., Lombard function has a slope ranging from $13/(90-50) \doteq 0.3$ dB/dB to $13/(90-60) \doteq 0.4$ dB/dB, approximately). This corresponds well with the slope values reported in past studies, (Lane *et al.*, 1961), (Webster and Klumpp, 1962), see Sec. 3.2.1 for details.

Parameter		Neutral		Noisy	
		Channel 0	Channel 1	Channel 0	Channel 1
Czech SPEECON	$\overline{\text{SNR}}$ (dB)	29.9	18.3	10.7	3.5
	σ_{SNR} (dB)	5.3	4.9	8.6	7.3
CZKCC	$\overline{\text{SNR}}$ (dB)	26.3	24.2	12.6	7.6
	σ_{SNR} (dB)	10.0	11.5	5.8	4.7
CLSD'05	$\overline{\text{SNR}}$ (dB)	27.9	10.9	40.9	28.8
	σ_{SNR} (dB)	7.2	5.2	6.8	7.1

Table 6.1: Means and deviations of SNR *distributions: Czech Speecon, CZKCC and CLSD'05. In CLSD'05, 'noisy' conditions refer to clean speech acquired in simulated noisy conditions.*

6.4 Fundamental Frequency

The F_0 tracking was performed separately for male and female speakers to evaluate differences in speech production across genders and recording conditions. F_0 distributions obtained for speech from neutral and noisy scenarios are shown in Fig. 6.2. F and M denote female and male data respectively. Distribution means and deviations are shown in Table 6.2. As could have been expected, female F_0

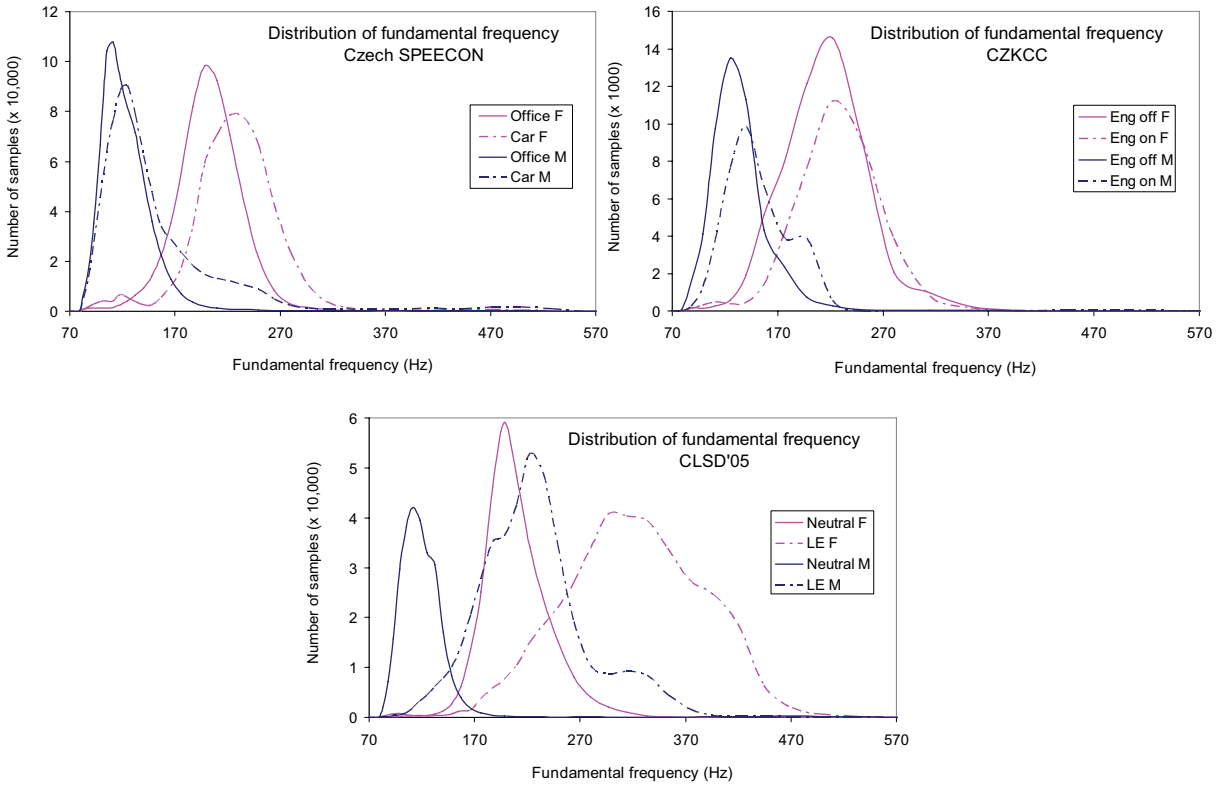

Figure 6.2: F_0 distributions: Czech Speecon, CZKCC and CLSD'05.

distributions are located at higher frequencies compared to male distributions. In all three databases, upward shifts of mean F_0 and increases of F_0 standard deviation can be observed when switching from neutral to noisy conditions, which confirms previous observations summarized in Sec. 3.2.2. In the case of CLSD'05, both the F_0 mean and standard deviation shifts are most considerable. Here, the

Parameter		Neutral		Noisy	
		Females	Males	Females	Males
Czech SPEECON	$\bar{F}_0$ (Hz)	197.7	121.1	220.1	140.3
	σ_{F0} (Hz)	33.6	27.5	46.5	51.8
CZKCC	$\bar{F}_0$ (Hz)	208.2	129.9	220.2	147.1
	σ_{F0} (Hz)	40.4	26.2	42.3	33.8
CLSD'05	$\bar{F}_0$ (Hz)	204.8	117.8	307.2	215.7
	σ_{F0} (Hz)	32.2	26.1	66.7	50.1

Table 6.2: F_0 means and standard deviations: Czech Speecon, CZKCC and CLSD'05.

mean male LE F_0 reaches a higher frequency than the mean female neutral F_0. The female LE F_0 distribution interferes with the typical location of F_1 appearance. Considering ASR tasks, the high F_0 values may cause a mismatch between feature vectors of the processed LE speech and acoustic models trained on neutral speech, resulting in a deterioration of recognition performance. In both genders, standard deviations of the LE F_0 distribution double compared to the neutral distribution.

6.5 Formants

Formant analysis was performed on the utterances containing digits. First, a monophone HTK recognizer, (Young *et al.*, 2000), see Sec. 4.7.1, was trained on 70 male and female SPEECON office sessions. The recognizer was used to perform *forced alignment* (i.e., to find phoneme boundaries in the speech signal given the known transcription of the phonetic content). Second, formant tracking was performed. Information about the first four formant frequencies and bandwidths was assigned to the corresponding phones. Since the formant values came from segments of different lengths, weighted means and deviances were calculated, see Sec. 4.6.1.

Mean locations of the first two formants in digit vowels are shown in Fig. 6.3. N refers to the neutral conditions and LE represents the noisy conditions. To outline distribution of the F_1, F_2 samples, 1–σ ellipses[2] covering 39.4 % of the data are also depicted in the figure. Note that surfaces of the ellipses are relatively large. This is presumably caused by the fact that the formant values were extracted from the whole phone segments, including transitions to the adjacent phones. To obtain more 'focused' estimations of vowel locations in the F_1, F_2 plane, instead of averaging all values from the segment, a sample from the middle of the segment can be picked as the representative estimate. Another approach would be to pick a median value from the segment values. However, since the presented analyses focus on evaluations of phenomena affecting ASR, and monophone models also deal with the transitions on the segment boundaries, author of the thesis preferred to take values averaged across the whole segment as the phoneme representatives.

As can be seen in Fig. 6.3, formant shifts in Czech SPEECON and CZKCC are not significant since the error ellipses almost completely overlap. The smallest shifts can be observed in CZKCC, where the same speakers produced speech in quiet and in noise. In the case of CLSD'05, considerable shifts of vowel formants were displayed. For all vowels, and upward shift in frequency can be seen for both F_1 and F_2 here. No consistent shifts of F_3 and F_4 were noticed for any of the databases.

Changes in formant bandwidths did not display any consistent trend in Czech SPEECON and CZKCC. In contrast to this, in CLSD'05, the mean bandwidths of all first four formants were systematically reduced for Lombard speech. Means and deviations of F_1 bandwidth in the digit vowels are shown in Table 6.3. The statistical significance of $F_{1,2}$ and $B_{1,2}$ shifts between neutral and noisy

[2]See Sec. 4.2.1

Czech SPEECON								
Vowel	B_{1M} (Hz)	σ_{1M} (Hz)	*B_{1M} (Hz)*	*σ_{1M} (Hz)*	B_{1F} (Hz)	σ_{1F} (Hz)	*B_{1F} (Hz)*	*σ_{1F} (Hz)*
/a/	270	93	*234*	*111*	271*	84	*259**	*119*
/e/	157	84	*199*	*129*	221	82	*202*	*132*
/i/	116	46	*182*	*109*	156	65	*186*	*123*
/o/	252*	107	*262**	*111*	306	91	*273*	*130*
/u/	149	78	*195*	*109*	181	65	*202*	*123*

CZKCC								
Vowel	B_{1M} (Hz)	σ_{1M} (Hz)	*B_{1M} (Hz)*	*σ_{1M} (Hz)*	B_{1F} (Hz)	σ_{1F} (Hz)	*B_{1F} (Hz)*	*σ_{1F} (Hz)*
/a/	207*	74	*210**	*84*	275	97	*299*	*78*
/e/	125*	70	*130**	*78*	156	68	*186*	*79*
/i/	124*	49	*127**	*44*	105	44	*136*	*53*
/o/	275	87	*222*	*67*	263*	85	*269**	*73*
/u/	187	100	*170*	*89*	174*	96	*187**	*101*

CLSD'05								
Vowel	B_{1M} (Hz)	σ_{1M} (Hz)	*B_{1M} (Hz)*	*σ_{1M} (Hz)*	B_{1F} (Hz)	σ_{1F} (Hz)	*B_{1F} (Hz)*	*σ_{1F} (Hz)*
/a/	269	88	*152*	*59*	232	85	*171*	*68*
/e/	168	94	*99*	*44*	169	73	*130*	*49*
/i/	125	53	*108*	*52*	132*	52	*133**	*58*
/o/	239	88	*157*	*81*	246	91	*158*	*62*
/u/	134*	67	*142**	*81*	209	95	*148*	*66*

Table 6.3: Formant bandwidths – digit vowels. Italic letters represent noisy data. In CLSD'05, 'noisy data' refers to clean speech acquired in simulated noisy conditions. Pairs of values with asterisk did not reach statistically significant difference at 95% confidence level.

SPEECON		Vowel				
		/a/	*/e/*	*/i/*	*/o/*	*/u/*
M	#N/#LE	215/303	632/861	680/910	152/191	306/391
	F_1	N	N	+	N	+
	F_2	+	N	+	N	+
	B_1	+	+	+	N	+
	B_2	+	+	+	+	+
F	#N/#LE	168/229	441/540	478/628	108/133	198/277
	F_1	N	N	+	N	+
	F_2	+	+	+	N	+
	B_1	N	+	+	+	+
	B_2	+	+	+	+	+

CZKCC		Vowel				
		/a/	*/e/*	*/i/*	*/o/*	*/u/*
M	#N/#LE	408/395	811/923	388/441	123/134	381/402
	F_1	+	+	+	N	N
	F_2	N	+	N	N	N
	B_1	N	N	N	+	+
	B_2	+	+	+	+	+
F	#N/#LE	455/437	895/823	418/427	148/155	396/357
	F_1	N	+	+	N	N
	F_2	+	+	N	N	N
	B_1	+	+	+	N	N
	B_2	+	+	+	+	+

CLSD'05		Vowel				
		/a/	*/e/*	*/i/*	*/o/*	*/u/*
M	#N/#LE	331/1645	861/3965	549/1796	134/524	272/1046
	F_1	+	+	+	+	+
	F_2	+	+	+	N	+
	B_1	+	+	+	+	N
	B_2	+	+	+	+	+
F	#N/#LE	1297/990	3144/2478	1433/1240	428/352	654/578
	F_1	+	+	+	+	+
	F_2	+	+	+	+	+
	B_1	+	+	N	+	+
	B_2	+	+	+	+	+

Table 6.4: Significance of feature shifts between neutral and noisy conditions. '+' – neutral/LE parameter pairs reaching statistically significant difference at 95% confidence level, 'N' – other pairs. In CLSD'05, 'noisy' conditions refer to clean speech acquired in simulated noisy conditions.

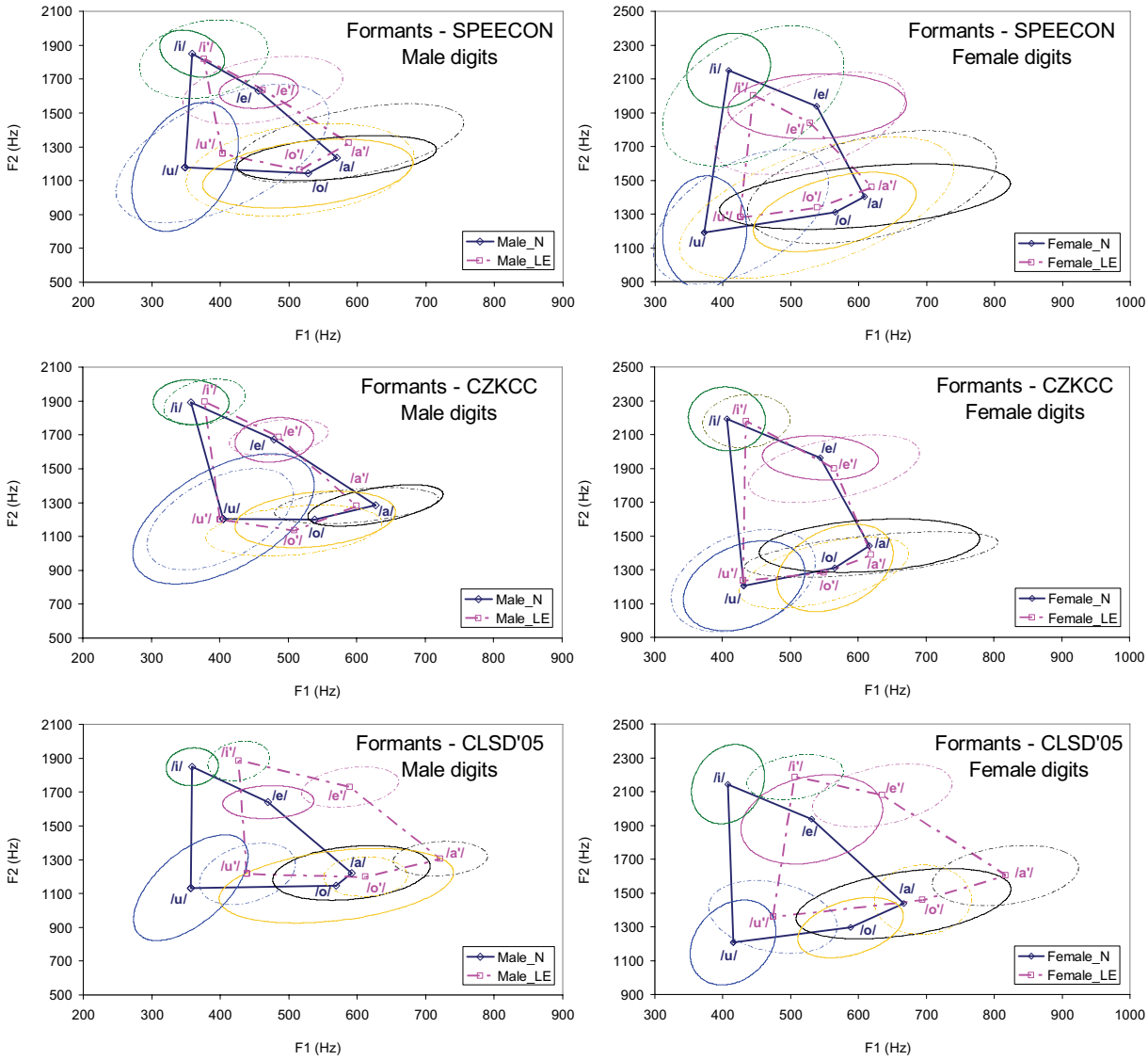

Figure 6.3: Formants F_1, F_2 – digit vowels.

data[3] is shown in Table 6.4. $\#N$ denotes the number of neutral realizations and $\#LE$ the number of noisy realizations of the phoneme parameter. The parameter pairs which displayed a statistically significant difference at a confidence level of 95 % are labeled '+', the rest is labeled 'N'. The lowest number of statistically significant shifts in features was displayed by CZKCC, while in CLSD'05, the majority of parameters changed significantly.

6.6 Durations

Average phoneme and word durations were evaluated for utterances containing digits, see Table 6.5. The durations were extracted from the phoneme boundaries obtained from the *forced alignment*, see Sec. 6.5. In Czech SPEECON, phoneme duration differences, see Sec. 4.5, did not exceed 38 %. In the case of CZKCC, the greatest duration changes were observed in the word 'štiri' (phoneme /r/

[3]In CLSD'05, 'noisy' data refer to clean speech acquired in simulated noisy conditions.

Czech SPEECON								
Word	Phoneme	# Office	T_{Office} (s)	$\sigma_{TOffice}$ (s)	# Car	T_{Car} (s)	σ_{TCar} (s)	Δ (%)
Štiri	*/r/*	135	0.040	0.022	233	0.052	0.041	29.87
***P**jet*	*/p/*	134	0.050	0.026	174	0.069	0.134	37.92
*Pj**e**t*	*/e/*	134	0.059	0.038	174	0.070	0.067	18.33*
*Dev**j**et*	*/v/*	129	0.049	0.016	174	0.067	0.145	35.52*

CZKCC								
Word	Phoneme	# OFF	T_{OFF} (s)	σ_{TOFF} (s)	# ON	T_{ON} (s)	σ_{TON} (s)	Δ (%)
*Nul**a***	*/a/*	349	0.147	0.079	326	0.259	0.289	48.50
*Jedn**a***	*/a/*	269	0.173	0.076	251	0.241	0.238	39.36
*Dv**a***	*/a/*	245	0.228	0.075	255	0.314	0.311	38.04
*Štir**i***	*/r/*	16	0.045	0.027	68	0.080	0.014	78.72
Sedm	*/e/*	78	0.099	0.038	66	0.172	0.142	72.58

CLSD'05								
Word	Phoneme	# N	T_N (s)	σ_{Tn} (s)	# LE	T_{LE} (s)	σ_{Tle} (s)	Δ (%)
Jedna	*/e/*	583	0.031	0.014	939	0.082	0.086	161.35
*Dvj**e***	*/e/*	586	0.087	0.055	976	0.196	0.120	126.98
*Čtir**i***	*/r/*	35	0.041	0.020	241	0.089	0.079	115.92
*Pj**e**t*	*/e/*	555	0.056	0.033	909	0.154	0.089	173.71
Sedm	*/e/*	358	0.080	0.038	583	0.179	0.136	122.46
Osm	*/o/*	310	0.086	0.027	305	0.203	0.159	135.25
*Devj**e**t*	*/e/*	609	0.043	0.022	932	0.120	0.088	177.20

Table 6.5: Phoneme durations. '' – pairs that did not reach statistically significant difference.*

– 79 %) and in the word 'sedm' (phoneme /e/ – 73 %). The most significant and consistent phoneme duration differences were observed in the CLSD'05 database. The highest differences can be found in the words 'devjet' (2^{nd} /e/ – 177 %), 'pjet' (/e/ – 174 %), and 'jedna' (/e/ – 161 %).

No significant changes in word durations were observed in Czech SPEECON. In CZKCC and CLSD'05, certain changes in word durations can be found, but do not reach the ratios of the phoneme changes, see Table 6.6. This results from the fact that when produced in noise, vowels tend to be increased in duration while the duration of consonants tends to be reduced (see Sec. 3.2.5).

CZKCC							
Word	# OFF	T_{OFF} (s)	σ_{OFF} (s)	# ON	T_{ON} (s)	σ_{TON} (s)	Δ (%)
Nula	349	0.475	0.117	326	0.560	0.345	17.82
Jedna	269	0.559	0.136	251	0.607	0.263	8.58
Dva	245	0.426	0.106	255	0.483	0.325	13.57

CLSD'05							
Word	# N	T_N (s)	σ_{Tn} (s)	# LE	T_{LE} (s)	σ_{Tle} (s)	Δ (%)
Nula	497	0.397	0.109	802	0.476	0.157	19.87
Jedna	583	0.441	0.128	939	0.527	0.165	19.56
Dvje	586	0.365	0.114	976	0.423	0.138	15.87

Table 6.6: Word durations.

6.7 Digit Recognition Task

The recognizer mentioned in Sec. 4.7.1 was also employed in the digit recognition task. The test set comprised isolated and connected digits. The resulting performances are shown in Table 6.7. The row

Set	*Czech SPEECON*				*CZKCC*				*CLSD'05*			
	Office F	Office M	Car F	Car M	OFF F	OFF M	ON F	ON M	N F	N M	LE F	LE M
# Spkrs	22	31	28	42	30	30	18	21	12	14	12	14
# Digits	880	1219	1101	1657	1480	1323	1439	1450	4930	1423	5360	6303
WER (%)	5.5 (4.0–7.0)	4.3 (3.1–5.4)	4.6 (3.4–5.9)	10.5 (9.0–12.0)	3.0 (2.1–3.8)	2.3 (1.5–3.1)	13.5 (11.7–15.2)	10.4 (8.8–12.0)	7.3 (6.6–8.0)	3.8 (2.8–4.8)	42.8 (41.5–44.1)	16.3 (15.4–17.2)

Table 6.7: Recognition performances: Czech SPEECON, CZKCC, and CLSD'05. Mean values followed by 95% confidence intervals in parentheses.

Set denotes type of scenario (*Offc – office*, *OFF – standing car/engine off*, *ON – moving car/engine on*, *N – neutral* conditions, *LE – simulated Lombard* conditions), and *WER* is the word error rate, see Sec. 4.7.3.

In Czech SPEECON, the error rate rose by 6 % for males under LE, with no significant change in *WER* observed for females. In CZKCC, the recognition performance was reduced by 8 % for male and by 11 % for female speakers. The most considerable deterioration of recognition performance was displayed on Czech CLSD'05, where *WER* increased by 12 % for male and by 36 % for female speakers.

6.8 Conclusions

Czech SPEECON, CZKCC, and CLSD'05 were used in feature analyses and digit recognition tasks. For all three databases, feature variations can be found when comparing utterances from quiet and noisy recording scenario. However, in the case of Czech SPEECON and CZKCC, shifts in feature distributions due to the speech production in noise are often negligible and inconsistent. It seems that the appearance of LE in the databases is only marginal. This may be a consequence of how the recording setup employed in the acquisition of the databases was defined. Speakers just read prompts from the list without the need to preserve speech intelligibility and react appropriately to the actual noise. This corresponds well with the observations referring to the 'lack of communication factor' in the databases as reported by the past works, see Sec. 3.6. Considering the minor changes in feature

distributions, the decreases in recognition performance on Czech SPEECON and CZKCC utterances in noisy conditions can be attributed more to the speech signal corruption by noise than to LE.

In CLSD'05, considerable shifts were found in the majority of the features analyzed. Since the recordings from both neutral and simulated noisy conditions contain speech signals with high *SNR*, it can be assumed that the feature tracking accuracy was less affected by the presence of noise compared to the other two databases. For the same reason, the significant drop in the recognition performance on the simulated LE utterances can be attributed exclusively to the feature shifts. The significantly stronger performance drop on the female utterances compared to the male utterances agrees with the observations reported by the earlier works, (Junqua *et al.*, 1998). Based on the results of feature analyses conducted in this chapter, it seems that some of the male features tend to shift towards the neutral female locations (F_0, F_1, F_2) while the female features shift to locations unseen in the neutral training data. Particularly, shifts of female F_0 distribution into the location of typical F_1 occurrence may strongly contribute to the performance corruption.

Results of the set of experiments carried out in this chapter demonstrate a strong presence of Lombard effect in the database. Hence, CLSD'05 was used in the experiments on Lombard speech conducted in the remainder of this thesis.

Chapter 7

Acoustic Model Adaptation

One of the possibilities to improve performance of an ASR system is to train its acoustic models on data coming from the actual environmental/talking style conditions, (Bou-Ghazale and Hansen, 1998), see also Sec. 3.5. This approach works well if a sufficient amount of training samples is available. However, if the conditions tend to change, it may be difficult to gather enough data for the model retraining on the fly. In such a case, acoustic model adaptation may be effective. Here, a relatively small amount of samples is used to adapt acoustic model parameters in order to match better the actual properties of the speech signal. In this chapter, Maximum Likelihood Linear Regression (MLLR) and Maximum A Posteriori approach (MAP) are used to adapt model means[1].

In MLLR, (Gales and Woodland, 1996), a set of transformations is applied to mean vectors of GMMs to map the former means $\boldsymbol{\mu}$ to the condition-dependent means $\boldsymbol{\mu}'$:

$$\boldsymbol{\mu}' = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \tag{7.1}$$

where $\boldsymbol{A}$ is the transformation matrix and $\boldsymbol{b}$ represents a bias vector. $\boldsymbol{A}$ and $\boldsymbol{b}$ are obtained by applying the expectation-maximization algorithm (EM)[2]. GMM variances may be adapted in a similar way, (Gales *et al.*, 1996).

If only a small amount of adaptation data is available, a *global adaptation transformation* is conducted. Here, a single transformation is used to adapt all GMMs. If more adaptation data becomes available, the number of transformations can be increased, yielding a *set of transformations*, each of them addressing a single group of GMMs. In this case, a dynamic method can be used to construct further transformations and to cluster groups of models as the amount of data increases, Young *et al.* (2000). The grouping of GMMs makes it possible to adapt distributions of acoustic models (e.g., some of the monophones or silences) that were not seen in the adaptation data. A regression class tree is built by clustering together components that are close in the acoustic space. The tree is designed based on the former (condition-independent) set of models, using a centroid splitting algorithm.

In the MAP adaptation approach, (Gauvain and Lee, 1994), the prior knowledge about the model parameter distribution is taken into account[3]. In adaptation of GMMs based on MAP, the condition-independent model parameters are used as informative priors, Young *et al.* (2000). The update formula

[1]The framework was provided by Petr Červa, Technical University of Liberec, who also conducted the model adaptation. Author of the thesis proposed the experiments, provided the data, and processed the results. The experiments were published in (Bořil *et al.*, 2006c).

[2]EM technique comprises two steps. In the *expectation* step, expectation of the likelihood is evaluated by initializing the latent variables of the model as if they were observed. In the *maximization* step, maximum likelihood estimates of the model parameters are found by maximizing the expected likelihood found in the *expectation* step. The two steps are usually repeated a couple of times. The parameters found in the *maximization* step are used to initialize the subsequent *expectation* step, (Wikipedia, 2007) – 'Expectation-Maximization Algorithm'.

[3]Let $\boldsymbol{\mu}_{MAP}$ be the parameter vector to be estimated. The adaptation can be performed using the maximum a posteriori estimate: $\boldsymbol{\mu}_{MAP} = \arg\max_{\boldsymbol{\mu}} p(\boldsymbol{o}|\boldsymbol{\mu})\, p(\boldsymbol{\mu})$, where $p(\boldsymbol{o}|\boldsymbol{\mu})$ is the density function for the adaptation data

is defined:

$$\boldsymbol{\mu}' = \frac{N}{N+\tau}\bar{\boldsymbol{\mu}} + \frac{\tau}{N+\tau}\boldsymbol{\mu}, \tag{7.2}$$

where N is the occupation likelihood of the adaptation data, τ is the weighting of the a priori knowledge, $\boldsymbol{\mu}$ is the condition-independent mean, and $\bar{\boldsymbol{\mu}}$ is the weighted mean of the adaptation data (the adaptation samples are weighted by their observation likelihoods given the condition-independent model).

7.1 Experiments

The model adaptation conducted in this chapter comprised two steps. First, mean vectors of GMMs in the speaker-independent (SI) models were transformed by MLLR. MLLR employed clustering of the models using a binary regression tree which allowed for adapting also models not represented sufficiently in the adaptation data set. Subsequently, the transformed values were used as the priors for the MAP-based adaptation. During MAP, transformations were performed only for the nodes where a sufficient amount of adaptation data was available.

When comparing genders in the baseline experiments, the recognition deterioration by LE was found considerably stronger for female speakers, see Sec. 6.7. For this reason, the presented experiments focused on improving the recognition performance on female utterances. Speaker-independent (SI)/gender-dependent (female) models of the LVCSR system, see Sec. 4.7.2, were adapted as follows:

- Speaker independent adaptation to Lombard speech: Neutral trained SI models are transformed to Lombard SI models – distinct utterances from the same or distinct speakers are used for the adaptation and for the open test, experiments 'SI adapt to LE (same speakers adapt/test)' or 'SI adapt to LE (different speakers adapt/test)', respectively.
- Speaker dependent (SD) adaptation to neutral speech: Neutral SI models were transformed into SD neutral models – 'SD adapt to neutral'.
- SD adaptation to Lombard speech: Neutral SI models were transformed into Lombard SD models – 'SD adapt to LE'.

All adaptation and recognition tasks were carried out twice, using either digit utterances (digit task) or phonetically rich sentences (sentences task), respectively, for adaptation and testing. Neutral and Lombard CLSD'05 sessions from 10 female speakers were used in the experiments. The data were partitioned to the adaptation set and to the open test as follows:

- SI adapt to LE (same speakers adapt/test): from all speakers, 2/3 of LE utterances were used for the SI model adaptation, the remaining 1/3 was used for the open test (880 digits, 970 words),
- SI adapt to LE (different speakers adapt/test): LE utterances from 6 speakers were used for SI model adaptation, the utterances from the remaining 4 speakers formed the test set (1024 digits, 1081 words),
- SD adapt to neutral: models are adapted to neutral utterances separately for each speaker, yielding speaker-dependent (SD) neutral models, and tested on LE speech in the SD recognition task (880 digits, 970 words),

sample $\boldsymbol{o}$ and $p(\boldsymbol{\mu})$ is the prior density of $\boldsymbol{\mu}$. If there is a prior knowledge about what are the model parameters likely to be – an *informative prior*, a limited amount of adaptation data may be sufficient for a reasonable MAP estimate. If there is no information about the prior density of $\boldsymbol{\mu}$ – *non-informative prior*, $p(\boldsymbol{\mu})$ becomes a uniform distribution and the resulting MAP estimate will be identical to the one obtained from the MLLR approach, (Dines, 2003), Young *et al.* (2000).

- SD adapt to LE: models are adapted to LE utterances separately for each speaker, yielding speaker-dependent (SD) LE models. From each LE session, 2/3 of utterances were used for the SD model adaptation, 1/3 for the SD open test (880 digits, 970 words).

The numbers in brackets refer to the number of digits used for the testing in the digit task or the number of words in the sentences task, respectively. In all experiments, the LVCSR system employed language model. Results of the experiments are shown in Table 7.1 and Fig. 7.1.

Digits	Models	Baseline	Baseline	Adapted
	Test set	Neutral	LE	LE
WER (%)	SI adapt to LE (same speakers train/test)	15.0 (12.6–17.4)	54.7 (51.4–58.0)	16.8 (14.3–19.3)
	SI adapt to LE (disjunct speakers train/test)	15.0* (12.6–17.4)	55.5 (52.5–58.5)	16.9 (14.6–19.2)
	SD adapt to neutral	15.0 (12.6–17.4)	54.7 (51.4–58.0)	43.9 (40.6–47.2)
	SD adapt to LE	15.0 (12.6–17.4)	54.7 (51.4–58.0)	8.5 (6.7–10.3)

Sentences	Models	Baseline	Baseline	Adapted
	Test set	Neutral	LE	LE
WER (%)	SI adapt to LE (same speakers train/test)	32.3 (29.4–35.2)	69.7 (66.8–72.6)	43.0 (39.9–46.1)
	SI adapt to LE (disjunct speakers train/test)	32.3* (29.4–35.5)	78.5 (76.1–80.9)	61.2 (58.3–64.1)
	SD adapt to neutral	32.3 (29.4–35.5)	69.7 (66.8–72.6)	68.7 (65.8–71.6)
	SD adapt to LE	32.3 (29.4–35.2)	69.7 (66.8–72.6)	39.2 (36.1–42.3)

Table 7.1: Efficiency of model adaptation: Digit and sentences tasks. '' – neutral open set together with neutral utterances from speakers participating in model adaptation. Mean values followed by 95% confidence intervals in parentheses.*

It can be seen that all adaptation setups improved the LVCSR performance. Speaker-dependent adaptation to LE provided the highest WER reduction, followed by speaker independent adaptations to LE and speaker dependent adaptation to neutral speech. The superior performance of speaker-dependent adaptation to LE is not so surprising as in this case, adaptation data allows for modeling both speaker-dependent and condition-dependent characteristics similar to the test data. Speaker-dependent adaptation to neutral speech was the least effective strategy.

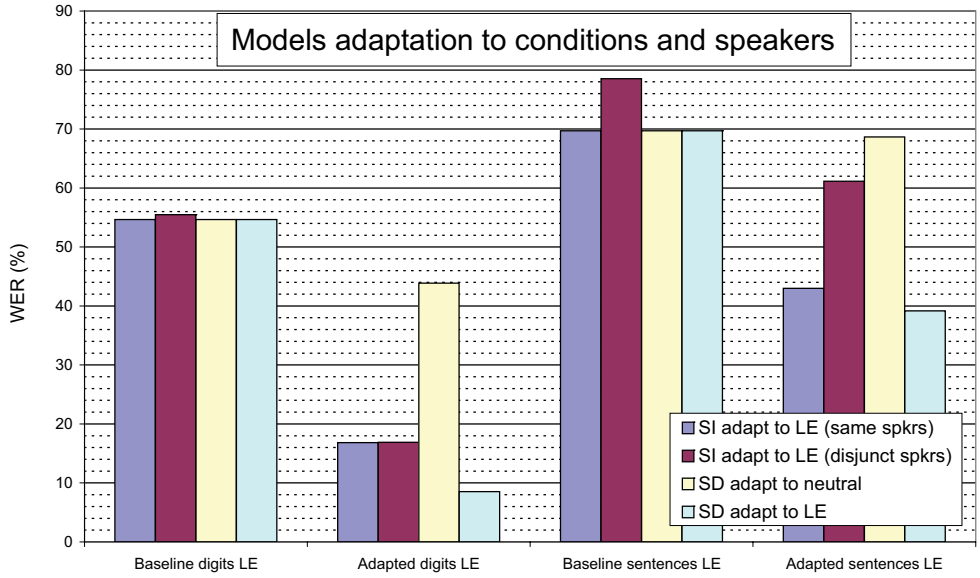

Figure 7.1: Overall performance of model adaptation.

7.2 Conclusions

All configurations of the acoustic model adaptation considered in this chapter improved recognition performance:

- The best results were reached by the SD adaptation to LE. In this case, the transformed models address both speaker-specific and LE-specific characteristics of speech provided in the adaptation data,

- In general, various speakers tend to react to the same noisy conditions differently, which may discourage using the speaker-independent adaptation to LE. However, a remarkable improvement was displayed by both experiments of SI adaptation to LE. A comparable recognition performance was reached when using distinct utterances from the same set of speakers for the adaptation and testing, and for the adaptation using distinct speakers for the model transformation and testing. In the first case, the adaptation is not pure SI but rather a group dependent (GD) task, as the same group of speakers participated in both model adjustments and the recognition task. In the case of the sentences task, the 'pure' SI scenario was less effective than for GD. This is presumably caused by the bigger complexity of the sentences task. Compared to the digits, here, an extensive number of units (phonemes, words) is considered during the recognition. Their distances in the acoustic space are smaller than in the case of digits, hence, a relatively small feature drift from the neutral model distributions may result in a considerable corruption of the recognition performance.

- SD adaptation to neutral speech was the least efficient approach. In the case of the sentences task, the recognition improvement was not statistically significant. This proves that SI models deal well with the speaker changes, but are strongly sensitive to the changes due to LE.

In real-world tasks, acoustic model adaptation can be applied to continuously update the ASR system parameters. In such a setup, incoming utterances would be first decoded and phone-aligned using the current acoustic models. Once a sufficient amount of data is available for a particular acoustic class or a class group, the corresponding acoustic models can be transformed towards the

actual speech parameters. Due to the low accuracy of the baseline neutral models when exposed to LE speech, the estimated transcriptions of the adaptation data can be expected to contain a lot of errors, which may slow-down or corrupt the convergence of the adapted models towards the actual speech characteristics.

In the experiments conducted in this chapter, only adaptation of mean vectors was conducted. Since LE is known to affect not only means but also variances of cepstral coefficients, see (Takizawa and Hamada, 1990), it can be assumed that further extension of the adaptation framework for variance transformation will yield improved modeling of the actual speech.

Chapter 8

Voice Conversion

Voice conversion is a technique originally developed for transforming speech from the source speaker towards a target speaker–sounding speech, (Abe *et al.*, 1988), (Childers, 1995), (Baudoin and Stylianou, 1996), (Kain and Macon, 1998), (Arslan, 1999). Since voice conversion typically addresses speech production differences between source and target speech both on the excitation and vocal tract level, it seems to be a promising means also for transforming talking styles. In past studies, voice conversion techniques employing a combination of speech perturbation models and a code-excited linear prediction (CELP), (Bou-Ghazale and Hansen, 1996), or HMM modeling, (Bou-Ghazale and Hansen, 1998), were used to generate simulated stressed speech from neutral speech. In the latter case, the obtained samples were employed in training a speech recognizer, significantly improving its performance on actual stressed (angry and loud) speech. In this chapter, a novel voice conversion-based approach to Lombard speech equalization is presented[1]. First, feature distributions in source Lombard utterances and target equalized utterances are analyzed. Second, efficiency of the LE equalization when being integrated into the front-end of the neutral-trained ASR system is evaluated in the digit and LVCSR task.

Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$ be a sequence of feature vectors extracted from the frames of the source speaker's speech signal and $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N$ be a sequence of the corresponding feature vectors extracted from the target speaker's speech signal. When training a voice conversion system, the goal is to find a conversion function F minimizing the mean square error

$$\varepsilon_{MSE} = E\left[\left\|\boldsymbol{y}_n - F\left(\boldsymbol{x}_n\right)\right\|^2\right], \tag{8.1}$$

where E denotes expectation and $\|\cdot\|$ is the Euclidean norm defined:

$$\|\boldsymbol{z}\| = \sqrt{z_1 + z_2 + \ldots + z_M}, \tag{8.2}$$

and M is the number of components of the vector $\boldsymbol{z}$.

Various approaches to voice conversion have been explored: quantization of source and target feature vectors and applying codebook transformations, (Abe *et al.*, 1988), warping short-time spectra of the source speaker towards the target speaker spectra, (Sündermann *et al.*, 2005c), continuous probabilistic spectral transformation based on joint density GMMs, (Kain and Macon, 1998), (Sündermann *et al.*, 2005b), and transformation of the vocal tract residuals, (Sündermann *et al.*, 2005a).

[1]The idea of using voice conversion for Lombard speech equalization was proposed by Prof. Harald Höge, Siemens Corporate Technology, Munich, Germany (Siemens CT), and investigated within the frame of the joint project 'Normalization of Lombard Effect' of Siemens CT and CTU in Prague, (Bořil, 2007). David Sündermann, Siemens CT, provided a voice conversion system and conducted its training and data conversion, (Sündermann *et al.*, 2005b). Author of the thesis provided parallel utterances for the system training, analyzed an impact of voice conversion on feature distributions, and evaluated efficiency of the conversion in the recognition tasks. Results of the experiments were presented in (Bořil *et al.*, 2006c).

Voice conversion algorithms usually employ *text-dependent training* which requires availability of parallel utterances from the source and target speaker. In this chapter, GMM-based *text-dependent* voice conversion is used. Here, the source speech signal is decomposed to the vocal tract and excitation components, see Sec. 4.4, which are then transformed separately. Training of the system comprises the following steps:

- Utterances with the same linguistic content (parallel utterances) from the source and target speaker are used for the training,
- The speech signals are split into pitch-synchronous frames. In unvoiced regions, the speech signal is segmented based on the interpolation of lengths of the neighboring voiced segments,
- The parallel source and target frame sequences are aligned using dynamic time warping (DTW), see Chap. 2,
- From the aligned sequences, *vocal tract-related features* are extracted and used for training the joint density GMM:

$$p\left(\boldsymbol{x},\boldsymbol{y}\right)=\sum_{m=1}^{M}c_m\mathcal{N}\left\{\left[\begin{array}{c}\boldsymbol{x}\\ \boldsymbol{y}\end{array}\right];\boldsymbol{\mu}_m=\left[\begin{array}{c}\boldsymbol{\mu}_m^x\\ \boldsymbol{\mu}_m^y\end{array}\right],\boldsymbol{\Sigma}_m=\left[\begin{array}{cc}\boldsymbol{\Sigma}_m^{xx} & \boldsymbol{\Sigma}_m^{xy}\\ \boldsymbol{\Sigma}_m^{yx} & \boldsymbol{\Sigma}_m^{yy}\end{array}\right]\right\}, \tag{8.3}$$

  where c_m is the weight of the m-th mixture component (see Sec. 2.2), $\boldsymbol{\mu}_m^x$ and $\boldsymbol{\mu}_m^y$ are mean vectors of the source and target features, and $\boldsymbol{\Sigma}_m$ is the covariance matrix. From the GMM, the linear transformation function minimizing the mean square error between the converted and target vocal tract features is derived, (Kain and Macon, 1998):

$$F_V\left(\boldsymbol{x}\right)=\sum_{m=1}^{M}p_m\left(\boldsymbol{x}\right)\left[\boldsymbol{\mu}_m^y+\boldsymbol{\Sigma}_m^{yx}\left(\boldsymbol{\Sigma}_m^{xx}\right)^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_m^x\right)\right], \tag{8.4}$$

  where $p_m\left(\boldsymbol{x}\right)$ is the probability that the feature vector $\boldsymbol{x}$ belongs to the m-th Gaussian component.

- Average and variance of the target pitch are also determined. During voice conversion, the pitch of the source *excitation* component F_{0x} is adjusted to match the target pitch F_{0y} parameters, (Kain and Macon, 1998):

$$F_E\left(F_{0x}\right)=\mu_{F_{0y}}+\frac{\sigma_{F_{0y}}}{\sigma_{F_{0x}}}\left(F_{0x}-\mu_{F_{0x}}\right). \tag{8.5}$$

It may be difficult to obtain parallel utterances from the source and target speakers, especially if the speakers do not speak the same language. This can be addressed by a *text-independent* voice conversion approach, (Sündermann *et al.*, 2006a). Here, the phonetically corresponding frames are found by means of minimizing the target and concatenation frame-mapping costs. The target cost measures distance between source and target feature vectors. The concatenation cost evaluates the connection quality of the consecutive units, taking into account the continuity of spectral evolution of the natural speech.

Parallel Lombard and neutral utterances from the same speakers were used for the *text-dependent* training of the *speaker-dependent* voice conversion. Mean F_0 of source and target utterances were derived from the mean pitch-synchronous frame lengths. Besides F_0, the remaining parameters of the target excitation signal were kept equal to the source excitation signal. Coefficients of the linear predictive models representing vocal tract properties were transformed to line spectrum frequencies (LSF), (Kleijn *et al.*, 2003). The source LSF features were converted to the target features, transformed back to the time domain, and concatenated by means of the pitch-synchronous overlap and add (PSOLA) technique, (Charpentier and Stella, 1986).

8.1 Converted Speech Features

To evaluate separately the efficiency of the excitation equalization and the vocal tract equalization, two variants of voice conversion were considered in the presented experiments. In the first case, both F_0 and formants were transformed towards neutral – 'converted LE' (CLE). In the second case, only F_0 was converted towards neutral while formant components were preserved – 'converted LE – F_0 only' ($CLEF_0$). For each speaker, a *speaker-dependent* voice conversion system was trained on 20 parallel Lombard/neutral phonetically rich sentences. For the open test set experiments, phonetically rich Lombard utterances from 11 male and 11 female speakers – 'sentences' task, and digit Lombard utterances from 11 male and 10 female speakers[2] – 'digit' task, were equalized by conducting the speaker-dependent conversion.

SNR distributions of the *neutral*, LE, CLE, and $CLEF_0$ data are depicted in Fig. 8.1. Corresponding distribution means and standard deviations are shown in Table 8.1. SNR distributions

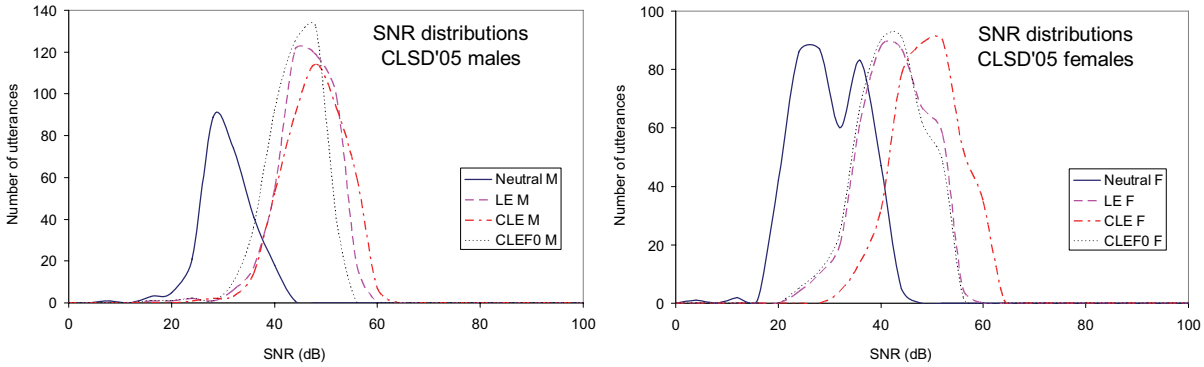

Figure 8.1: SNR distributions: neutral, LE, CLE, and $CLEF_0$ data; digits + sentences.

Set		Neutral	LE	CLE	$CLEF_0$
Males	$\overline{SNR}$ (dB)	30.7	46.2	47.7	43.9
	σ_{SNR} (dB)	5.0	5.6	5.9	5.3
Females	$\overline{SNR}$ (dB)	29.9	42.7	49.1	42.0
	σ_{SNR} (dB)	6.6	6.5	6.4	6.5

Table 8.1: SNR distribution means and standard deviations: neutral, LE, CLE, and $CLEF_0$ data; digits + sentences.

of the converted speech reach high mean values, occurring in the similar or close locations to the LE distributions since the voice conversion considered in this chapter preserved intensity as seen in the source speech. Hence, the impact of background noise on the accuracy of feature analyses and on the performance in recognition tasks can be assumed negligible in the case of the converted speech.

Distributions of F_0 and their means and standard deviations are shown in Fig. 8.2 and Table 8.2. It can be seen that F_0 distributions of the converted speech are shifted down the frequency axis and almost completely overlap the neutral distributions. The standard deviations of converted F_0 are also reduced towards the neutral values, although here, the transformation is less effective. Differences between means and deviations of the converted and neutral distributions may be partly attributed to the fact that the speaker-dependent conversion parameters were derived from a relatively small set

[2]One of the 11 female neutral sessions did not contain a complete set of parallel digit utterances, hence, only 10 complete female sessions were used.

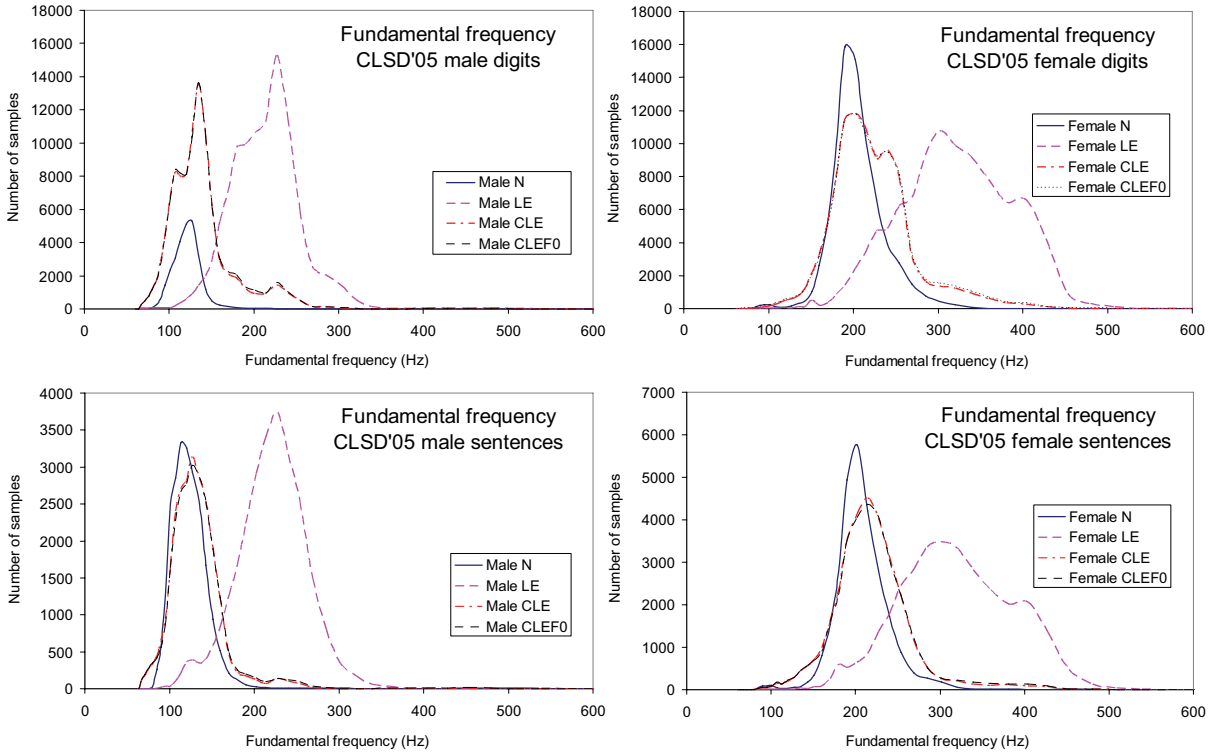

Figure 8.2: F_0 distributions: neutral, LE, CLE, $CLEF_0$.

of data disjunct to the open set. When summarizing results across the entire speaker set, it seems that the converted F_0 distributions display consistent shift compared to the neutral speech. Further explanation of this effect could be presumably derived from the particular properties of the algorithms used in the voice conversion system, which are not known to the author of the presented thesis.

Mean locations of the first two vowel formants F_1, F_2 are shown in Fig. 8.3. In the case of $CLEF_0$, only the F_0 component was transformed while the formant structure was kept intact. This corresponds well with the analysis results for higher formant values, where F_1, F_2 reach similar locations for $CLEF_0$ and LE vowels. In the case of the lowest F_1 values, typically in the vowels /u/ and /i/, a considerable difference between converted and LE formant locations can be found at times. This difference is presumably caused by occasional confusion of F_0 and F_1 in the automatic formant tracking algorithm. In most cases, CLE formants displayed shift towards the neutral locations. The only exception was male sentences where F_1 frequencies remained almost unchanged compared to the source LE data and mean F_2 values were frequently shifted to even further distances from the neutral vowels than the former LE data. In this case, a significant distortion of the speech signal was noticed in the listening tests. This distortion might be a reason for the failure of the automatic formant tracking algorithm.

Statistical significance[3] of the vowel F_1, F_2 location differences between the matched sets was analyzed, with results in Table 8.3 and Table 8.4. Features that did not display a significant shift between the two given sets are labeled N. The best joint performance of the voice conversion system and the formant tracking algorithm was reached for the male sentence pair LE–$CLEF_0$, where F_1, F_2 did not display any statistically significant shift after the F_0 conversion. Female sentence LE–$CLEF_0$ and male digit LE–$CLEF_0$ pairs also reported on an effective F_0 conversion. In other cases, the conversion was not accurate enough to result in consistent and statistically insignificant differences of the target and converted formants.

[3]See Sec. 4.6.2.

Set		Neutral	LE	CLE	$CLEF_0$
Male digits	$\bar{F}_0$ (Hz)	119.3	207.2	131.1	131.8
	σ_{F0} (Hz)	17.8	40.4	35.4	35.9
Female digits	$\bar{F}_0$ (Hz)	201.1	307.5	213.8	214.7
	σ_{F0} (Hz)	31.9	67.1	46.1	47.9
Male sentences	$\bar{F}_0$ (Hz)	120.9	213.6	126.4	127.8
	σ_{F0} (Hz)	23.2	42.7	28.6	32.0
Female sentences	$\bar{F}_0$ (Hz)	203.2	305.1	210.9	211.9
	σ_{F0} (Hz)	32.0	67.9	44.3	46.9

Table 8.2: F_0 distribution means and standard deviations: neutral, LE, CLE, $CLEF_0$.

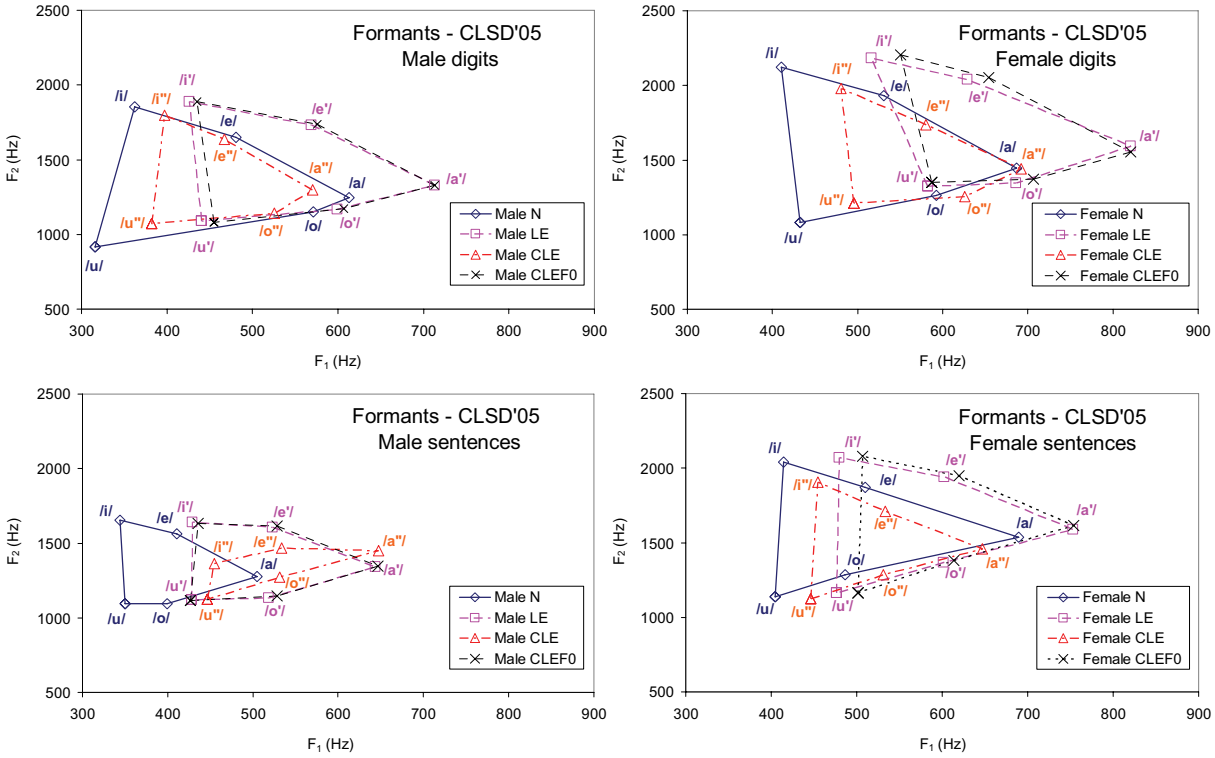

Figure 8.3: Vowel formants: neutral, LE, CLE, $CLEF_0$.

The F_1, F_2 bandwidths are shown in Table 8.5. Pairs labeled by asterisk represent values of statistically insignificant difference. As already discussed in Sec. 6.5, LE causes narrowing of the formant bandwidths. It can be seen that CLE compensates for this effect by widening the bandwidths, often to values exceeding the neutral bandwidths. Surprisingly, $CLEF_0$ bandwidths are consistently wider compared to LE bandwidths. Similarly as in the case of the F_0 conversion mismatch, this effect can be explained only based on the knowledge of the particular properties of the voice conversion system.

Parameter			Vowel				
			/a/	*/e/*	*/i/*	*/o/*	*/u/*
M	LE – $CLEF_0$	F_1	N	+	+	N	+
		F_2	N	N	N	N	N
	N – CLE	F_1	+	+	+	+	+
		F_2	+	+	+	N	+
F	LE – $CLEF_0$	F_1	N	+	+	+	N
		F_2	+	+	+	N	N
	N – CLE	F_1	N	+	+	+	+
		F_2	N	+	+	N	+

Table 8.3: Significance of formant shifts – digits. '+' – neutral/LE parameter pairs reaching statistically significant difference at 95% confidence level, 'N' – other pairs.

Parameter			Vowel				
			/a/	*/e/*	*/i/*	*/o/*	*/u/*
M	LE – $CLEF_0$	F_1	N	N	N	N	N
		F_2	N	N	N	N	N
	N – CLE	F_1	+	+	+	+	+
		F_2	+	+	+	+	N
F	LE – $CLEF_0$	F_1	N	+	+	N	N
		F_2	N	N	N	N	N
	N – CLE	F_1	+	+	+	+	+
		F_2	+	+	+	N	N

Table 8.4: Significance of formant shifts – sentences. '+' – neutral/LE parameter pairs reaching statistically significant difference at 95% confidence level, 'N' – other pairs.

Set	B_1 (Hz)				B_2 (Hz)			
	Neutral	LE	CLE	$CLEF_0$	Neutral	LE	CLE	$CLEF_0$
Male digits	297	136	209	156	268	169	284	208
Female digits	230*	164	223*	200	252	201	320	258
Male sentences	227	148	262	164	226	174	369	202
Female sentences	236	176	262	202	262	213	369	258

Table 8.5: Voice conversion – formant bandwidths. '' – pairs that did not reach statistically significant difference.*

8.2 Digit Recognition Task

Efficiency of the voice conversion–based LE normalization was evaluated in the digit and LVCSR task. In the digit task, the recognizer described in Sec. 4.7.1 employing gender independent models was used. Results of the experiment are shown in Table 8.6 and Fig. 8.4. In the case of male

Set	Male digits				Female digits			
	Neutral	LE	CLE	$CLEF_0$	Neutral	LE	CLE	$CLEF_0$
# Spkrs	11	11	11	11	10	10	10	10
# Digits	875	2816	2816	2816	2560	2560	2560	2560
WER (%)	2.3 (1.3–3.3)	16.5 (15.1–17.9)	17.3 (15.9–18.7)	13.7 (12.4–14.9)	4.2 (3.4–5.0)	43.6 (41.6–45.5)	25.4 (23.7–27.1)	34.9 (33.1–36.8)

Table 8.6: Voice conversion efficiency – digit recognition task. Mean values followed by 95% confidence intervals in parentheses.

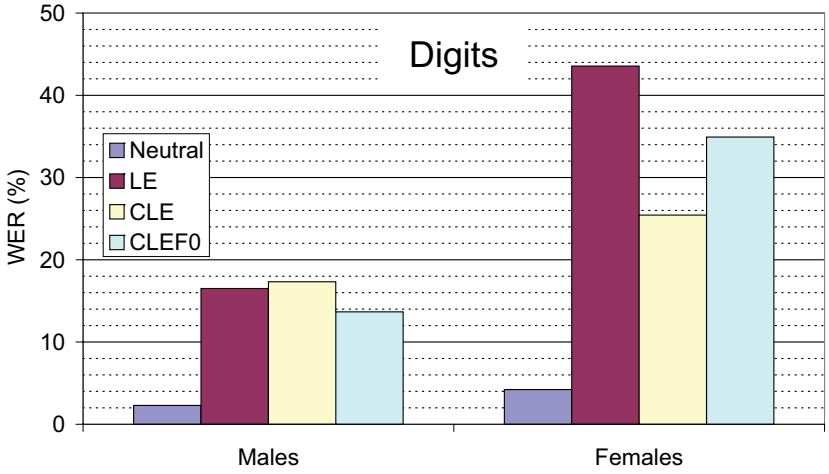

Figure 8.4: Voice conversion efficiency – digit recognition task.

speakers, a statistically significant improvement of the recognition performance was reached when applying $CLEF_0$. As shown in the previous section, in male speakers, $CLEF_0$ successfully equalized F_0 towards neutral while preserving the formant structure. Shifting F_0 to the lower frequencies helps to eliminate its interference with the location of typical F_1 occurrence, which might be otherwise confusing for the acoustic models of the recognizer. Applying CLE resulted in the performance drop, presumably due to the inaccuracy of the formant transformation discussed in the previous section.

In the case of female speakers, both CLE and $CLEF_0$ increased the recognition performance. Here, CLE displayed considerably better results. As already discussed in Sec. 6.5, while male LE F_1, F_2 tended to move to the female formant locations, female F_1, F_2 shifted to locations that were not covered in the neutral data used for the acoustic model training. Hence, even relatively inaccurate formant transformation towards neutral values was found helpful here.

8.3 LVCSR Task

In the LVCSR task, the recognizer described in Sec. 4.7.2, employing gender-dependent (GD) models, was used[4]. Two configurations of the recognizer were considered – a setup comprising language model (LM) and a setup with LM excluded. In the latter case, LM was substituted by the uniform word transition probabilities. About 3 % of the words occurring in the phonetically rich sentences were not covered in the vocabulary of the recognizer. Results of the experiment are shown in Table 8.7 and Fig. 8.5. Employing LM – 'GD/LM' – improved recognition performance for all data sets. In the

Set		Male sentences				Female sentences			
		Neutral	LE	CLE	$CLEF_0$	Neutral	LE	CLE	$CLEF_0$
# Spkrs		11	11	11	11	11	11	11	11
# Words		973	973	973	973	970	970	970	970
WER (%)	GD/no LM	77.9 (75.3–80.5)	85.3 (83.1–87.5)	91.5 (89.7–93.3)	85.5 (83.3–87.7)	72.9 (70.1–75.7)	86.5 (84.3–88.7)	90.1 (88.2–92.0)	87.9 (85.8–90.0)
	GD/LM	40.4 (37.3–43.5)	55.8 (52.7–58.9)	69.4 (66.5–72.3)	60.0 (56.9–63.1)	28.9 (26.0–31.8)	63.7 (60.7–66.7)	66.7 (63.7–70.0)	60.4 (57.3–63.5)

Table 8.7: Voice conversion efficiency – LVCSR task. Mean values followed by 95% confidence intervals in parentheses.

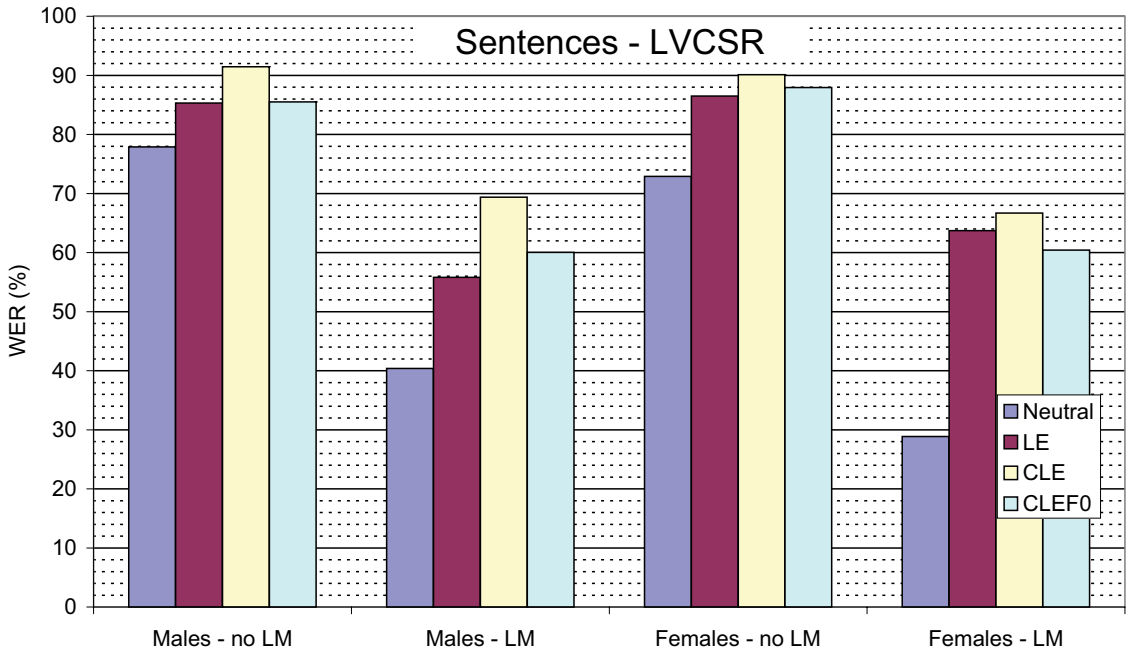

Figure 8.5: Voice conversion efficiency – LVCSR task.

neutral task, considerably lower WER was reached in the case of female utterances. This observation is consistent with the results of preliminary TUL experiments, where female acoustic models displayed consistently better performance than male models. Similarly as in the digit task, on the LE speech,

[4]The recognition task was conducted by Dr. Jindřich Žďánský, Technical University of Liberec. Author of the thesis proposed the experiment, provided data, and evaluated the results.

the degradation of the recognition accuracy was more significant in the female set, compare to Table 8.6. Voice conversion was not effective in the LVCSR task. Applying $CLEF$ corrupted recognition on all sets. $CLEF_0$ slightly improved the accuracy on the female sentences but was not successful on the male ones.

8.4 Conclusions

Speaker-dependent voice conversion employing *text-dependent training* was used to equalize LE speech towards neutral. Features of the source LE speech, converted speech, and target neutral speech were analyzed. Results of the analyses were affected by the joint contribution of the actual feature distributions and the accuracy of the feature tracking algorithms. In the case of LE and converted speech, features might have reached values unexpected by the ad hoc criteria employed in the tracking algorithms, resulting in analysis errors. For example, in LE speech, F_0 often reached values interfering with the location of the F_1 typical occurrence (see Fig. 8.2), which might have corrupted F_1 tracking.

In most cases, means and variances of F_0 and first two formants $F_{1,2}$ in the converted speech were transformed towards the neutral distributions. The least successful case of the tandem 'voice conversion/feature tracking' was found in male sentences, where a severe distortion of the formant structure occurred. In this case, when performing listening tests, significant distortion of the acoustic waveforms, often resulting in unintelligible speech, was observed.

In the digit recognition task, $CLEF_0$ improved recognition accuracy both on the male and female sets. State-of-the-art ASR systems are considered resistant to the variations of F_0. However, in LE speech, the F_0 shifts to the typical locations of F_1 may impact the ASR performance. This hypothesis is proven by the performance gains provided by $CLEF_0$ on LE speech, as here, only F_0 was transformed towards neutral while formants were preserved intact.

$CLEF$ was not effective on the male digit set. It seems that in this case, the effect of LE equalization was overruled by the inaccuracies in the formant conversion. In the case of female digits, $CLEF$ outperformed $CLEF_0$. Here, LE $F_{1,2}$ occurred in locations never seen by the recognizer during the acoustic model training and the formant transformation towards neutral, even though relatively inaccurate, was helpful.

In the LVCSR task, with one exception, the voice conversion–based equalization of LE was not effective. Compared to small vocabulary tasks, here, the words to be classified often occur very close in the acoustic space. Even a slight inaccuracy in feature equalization results in a severe deterioration of recognition performance.

As demonstrated on the digit task presented in this chapter, transformation of Lombard speech features towards neutral by means of voice conversion may be a promising way to improve performance of neutral ASR systems in LE. To allow for application of voice conversion–based LE equalization also in LVCSR tasks, more accurate transformation algorithms have to be defined first.

Chapter 9

Data-Driven Design of Robust Features

As previously discussed in Sec. 2.5, the goal of feature extraction in ASR is to provide a speech signal representation of reduced dimensionality, preserving linguistic information and suppressing variability introduced by speakers, environment, and signal processing chain. Various approaches to robust feature extraction were mentioned in Sec. 3.5. The majority of current ASR systems employ MFCC (Davis and Mermelstein, 1980) or, to a lesser degree, PLP (Hermansky, 1990) features. One of the key processing stages common to both algorithms is the smoothing of the FFT spectrum with a bank of nonlinearly distributed filters, see Sec. 2.5. Their distribution is derived from auditory models in an effort to emphasize the speech components essential for human speech perception. Some studies have reached front-end performance improvements by further modifying auditory based filter banks (FBs) (e.g., in Human Factor Cepstral Coefficients (HFCC) by changing bandwidths of mel filters, (Skowronski and Harris, 2004)). Others proposed new auditory models (e.g., Seneff auditory model comprising 40 filters matching a cat's basilar membrane response, (Seneff, 1986), or Ensemble Interval Histogram (EIH) model employing a bank of level crossing intervals, (Ghitza, 1988)). Also non-auditory, data-driven concepts of FB design were studied (e.g., Discriminative Feature Extraction method (DFE) iteratively adapting FB parameters, (Biem and Katagiri, 1997), or a design of a library of phoneme class-dependent filter banks, (Kinnunen, 2002)). Some of the filter banks introduced in these studies were tested in simulated noisy conditions, yet no extensive research on robustness to changes in talking style has been reported. In (Jankowski *et al.*, 1995), efficiency of various FBs for the processing of simulated loud utterances was evaluated, though not all properties of real loud speech were considered (e.g., F_0 and formant shifts). Suitability of various features including alternative filter bank processing, pre-emphasis, and cepstral mean normalization were studied for recognition of speech comprising different talking styles including LE, (Bou-Ghazale and Hansen, 2000). In that study, stressed speech recognition was enhanced by mel FB adjustments suppressing spectral mismatch between neutral models and stressed speech.

In this chapter, new LE-robust features employing filter banks obtained in the data-driven design are proposed[1] and compared to the selected set of standard and robust features used in recent ASR systems. Efficiency of the discrete cosine transform (DCT) and linear predictive coding (LPC) in the modeling of the short time amplitude spectra is compared. Performance of the features is evaluated from the view point of the resistance to changes of talking style (neutral/Lombard speech), average utterance pitch, and type and level of noisy background. Since the previous experiments have shown a significantly stronger corruption of recognition in case of female Lombard speech, see Sec. 8.2, 8.3, the presented feature design focused on female speakers. It can be presumed that the similar design scheme can be adopted successfully also for male speech.

[1]The feature design was published in (Bořil *et al.*, 2006b).

9.1 Development Setup

The recognizer described in Sec. 4.7.1 was used for the experiments presented in this chapter. HMM-based recognizers were trained on neutral female utterances exclusively (*train* set), yielding gender dependent models. The test data were split into *devel* (development) and *open* neutral and LE sets. The *devel* sets were used for the optimization of the front-end performance and the *open* sets for the open tests. Content of the data sets was as follows:

- *Train*: Czech SPEECON, 10 hours of signal, 37 female speakers, office sessions,
- *Devel* neutral: CLSD'05, 3480 words, 8 female speakers,
- *Devel* LE: CLSD'05, 3480 words, 8 female speakers,
- *Open* neutral: CLSD'05, 1450 words, 4 female speakers,
- *Open* LE: CLSD'05, 1880 words, 4 female speakers.

All data were down-sampled from 16 kHz to 8 kHz by *SoX*, (Norskog, 2007), and filtered by G.712 telephone filter using *FaNT* tool, (Hirsch, 2005). The joint transfer function of the decimation and telephone filter is shown in Fig. 9.1.

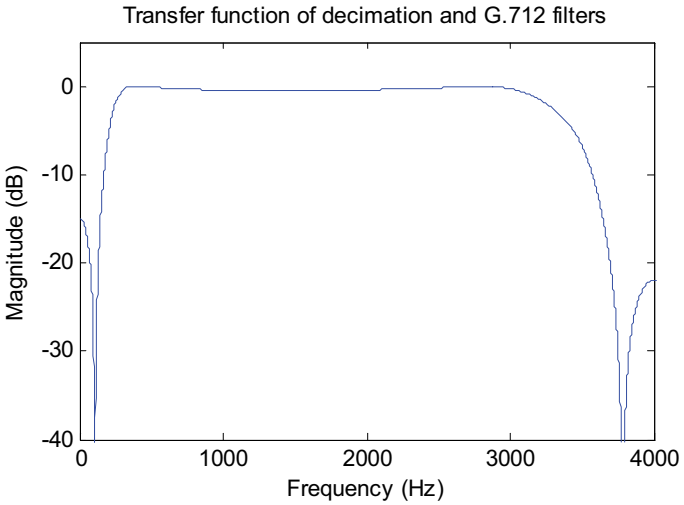

Figure 9.1: Joint transfer function of anti-aliasing decimation filter and G.712 telephone filter.

9.2 Baseline Features

In the initial experiment, performance of MFCC, PLP, and MR–RASTA when being employed in the front-end of monophone digit recognizer was evaluated.

9.2.1 MFCC and PLP

Feature extraction stages of MFCC and PLP were discussed in Sec. 2.5. Setup of the MFCC front-end is described in Sec. 4.7.1. PLP comprised a filter bank of 15 trapezoid filters and 12^{th} order LPC analysis. Similar signal segmentation and pre-emphasis as in the MFCC front-end were applied.

9.2.2 Multi-Resolution RASTA

Multi-resolution RASTA features (MR-RASTA) (Hermansky and Fousek, 2005) were extracted in 2 stages. First, an auditory spectrum with 15 bands was calculated from the speech similarly as in the PLP front-end[2]. Time trajectory of these sub-band energies was filtered with a bank of two-dimensional filters, yielding a set of about 500 coefficients every 10 ms. In the second step, an artificial neural network (ANN) projected the coefficients to the posterior probabilities of phones, reducing the feature vector size. The posteriors were then decorrelated and gaussianized using logarithm and principal component analysis in order to better fit the subsequent GMM-based HMM model[3].

9.2.3 Performance in Digit Recognition Task

Performance of MFCC, PLP, and MR-RASTA on the *devel* and *open* sets is summarized in Table 9.1. Neutral sets established a baseline at about 4 % WER. On LE data, a considerable decrease in accuracy was observed for all features. MFCC displayed the worst results, MR-RASTA significantly outperformed both MFCC and PLP. *Open* LE set seems to comprise more adverse data than *devel* LE set, since performance of all systems is reduced here.

Set		Devel set		Open set	
		Neutral	LE	Neutral	LE
# Digits		3480	3480	1335	1880
WER (%)	MFCC	3.6 (3.0–4.2)	63.3 (61.7–64.9)	3.7 (2.7–4.7)	68.7 (66.6–70.8)
	PLP	3.8 (3.2–4.4)	54.1 (52.4–55.8)	3.4 (2.4–4.4)	61.3 (59.1–63.5)
	MR-RASTA	4.5 (3.8–5.2)	39.8 (38.2–41.4)	4.1 (3.0–5.2)	42.1 (39.9–44.3)

Table 9.1: Performance of baseline features on female neutral and LE speech. Mean values followed by 95% confidence intervals in parentheses.

9.3 Designing Filter Banks

In (Arslan and Hansen, 1997), a relative significance of formant frequencies in the discrimination of accent and speech recognition was studied. The frequency interval 0–4 kHz was divided into 16 uniformly spaced frequency bands. The energy of each band was weighted by a triangular window and used as a single parameter for the training of HMMs for the accent classification and speech recognition. When evaluating performance of the single band-trained HMMs, it was observed that the impact of high frequencies on both speech recognition and accent classification performance was reduced[4]. The band 1500–2500 Hz representing interval of F_2–F_3 occurrence contributed most to accent classification while the lower frequency band 500–1700 Hz (F_1–F_2) was found the most important for speech recognition. A similar approach was exploited in (Bou-Ghazale and Hansen, 2000) to study how individual frequency bands are affected by the talking style changes. Here, a neutral,

[2]All features mentioned in this chapter were extracted using the open source tool *CtuCopy* developed by Dr. Petr Fousek, CTU in Prague, (Fousek, 2007).

[3]Training of the system employing MR–RASTA front-end was carried out by Dr. Petr Fousek, CTU in Prague.

[4]A similar concept can be found in (Steeneken and Houtgast, 1980), (Steeneken and Houtgast, 1999), where an amount of information content preservation within frequency bands is analyzed for the purposes of the objective measurement of speech intelligibility.

speaker-independent HMM recognizer was trained successively for each of the 16 bands and tested on neutral and angry speech. The highest recognition performance for neutral speech was reached using filters from the interval 200–1000 Hz, relating to the approximate location of F_1 occurrence. For angry speech, the interval 1250–1750 Hz (F_2) yielded the best results. Based on these observations, a new frequency scale, *Expolog*, increasing resolution in the interval of $F_{1,2}$ occurrence was proposed:

$$\mathrm{Expolog}\,(f) = \begin{cases} 700 \cdot \left(10^{\frac{f}{3988}} - 1\right), & 0 \leqslant f \leqslant 2000 \text{ Hz}, \\ 2595 \cdot \log\left(1 + \frac{f}{700}\right), & 2000 < f \leqslant 4000 \text{ Hz}. \end{cases} \tag{9.1}$$

Features employing FB of triangular filters distributed equidistantly on the *Expolog* scale outperformed MFCC on actual stressed noisy speech and on Lombard speech. The example of *Expolog* features[5] shows that knowledge of spectral distribution of linguistic information in the speech signal may provide a useful guideline for the design of a robust FB.

In *Expolog*, importance of a given frequency band was estimated based on the performance of the recognizer trained on this band, discarding information about the adjacent bands, overall spectral envelope, and inter-band dependencies (formant distances and bandwidths). However, a majority of the state-of-the-art features for ASR are extracted from the short-time spectral envelope. Hence, inspired by the *Expolog* design, a new approach to finding the linguistic *information distribution* (ID) is proposed in the following sections. The approach takes into account the nature of feature extraction by considering the impact of overall spectral envelope on ID. Based on the ID curves obtained from the development (*devel*) set, LE-robust FBs are designed.

9.3.1 Importance of Frequency Bands

The newly proposed method estimates a score-based ID across frequency bands by keeping all filters in FB but the examined one. Cepstral coefficients are extracted from the output energies of the remaining filters. In general, omitting one band can either increase the score compared to the baseline, meaning that the excluded band is dominated by irrelevant information, or decrease the score proving the excluded band's importance for the recognition.

The initial FB was chosen to consist of 20 linearly spaced rectangular filters without overlap (each of bandwidth 200 Hz), covering the frequency interval 0–4 kHz. No filter overlap was allowed so that omitting a FB filter assured that no component from the corresponding frequency band would contribute to the actual spectral envelope. The filters had rectangular shapes to provide uniform weighting of the frequency components being averaged in the filter outputs[6]. Besides the altered FB, the feature extraction comprised the same stages as standard MFCC.

ID curves obtained for neutral and Lombard *devel* sets are shown in Fig. 9.2. Baseline scores (no filter omitted from FB), are represented by dashed lines. For LE speech, the baseline score reached WER 29.0 %, significantly outperforming features presented in the previous section. Omitting the first FB band brings a slight degradation for neutral speech, but greatly enhances LE recognition (see also the first row in Table 9.2). In the case of LE speech, a significant peak in the ID curve can be observed in the region 600–1400 Hz (bands 4–7), interfering with the area of $F_{1,2}$ occurrence. For neutral speech, a corresponding peak lies in 400–1000 Hz (bands 3–5), the area of typical F_1 occurrence. This approximately agrees with the conclusions drawn for angry and neutral speech in (Bou-Ghazale and Hansen, 2000). Fig. 9.2 also suggests that Lombard speech recognition may be improved by avoiding low-frequency components at the expense of neutral speech recognition accuracy.

[5] Performance of *Expolog* features will be evaluated in Sec. 9.4.

[6] In (Psutka *et al.*, 2001), an influence of FB filter shapes on the front-end performance was studied, comparing rectangular, narrow trapezoid, wide trapezoid, and triangle filters. It was observed that filter shape had a negligible effect on the performance of the ASR system.

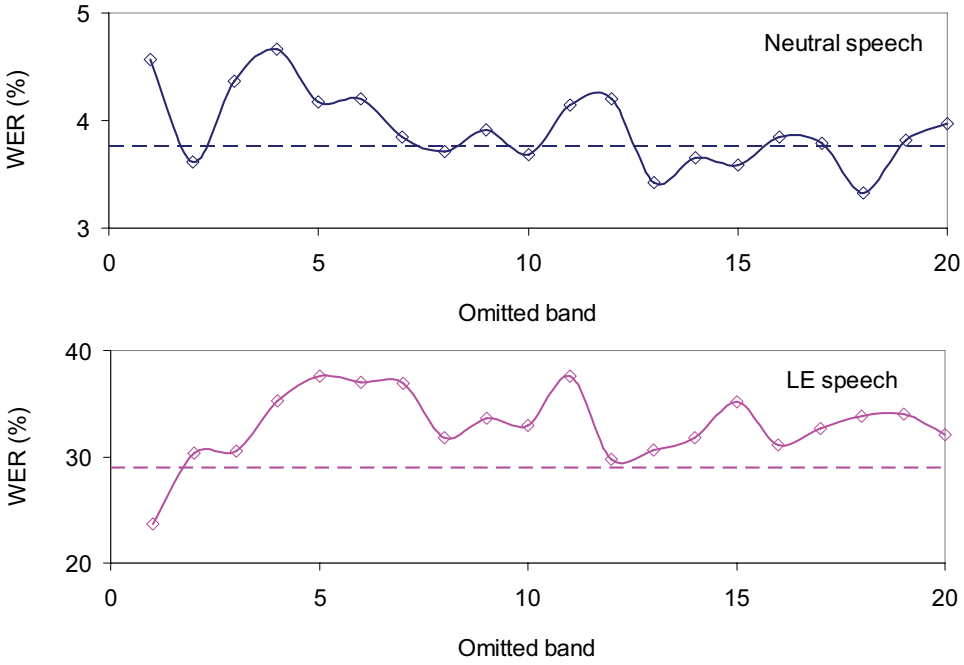

Figure 9.2: ID curves: impact of one missing band in 20–band FB on recognition performance.

A similar experiment was carried out on 43–bands FB covering 0–4 kHz but the ID obtained was noisy, not displaying any consistent trend of bands' importance. In this case, the omitted bands were presumably too narrow to noticeably affect the information content.

9.3.2 Avoiding Low Frequency Components

The previous section reported on an improvement trade-off between neutral and Lombard speech when avoiding low-frequency components. As a number of features efficient for neutral speech are available, the following design steps focused exclusively on LE recognition improvement.

As suggested by the previous section, the first band was excluded from the 20–bands FB. Subsequently, a dependency between the 19–bands FB's low cut-off frequency and recognition score was explored. The FB was first widened to cover the whole interval 0–4 kHz. In the subsequent steps, the FB's low cut-off frequency was gradually increased, while the high cut-off was kept constant. Bandwidths of all FB's filters were reduced in the same ratio to preserve uniform band distribution. The resulting WER/cut-off dependency is shown in Fig. 9.3. The minimum WER on Lombard speech was found at 625 Hz. WERs for the initial 20-bands FB and 19-bands FB starting at 625 Hz are shown in Table 9.2. Shifting the low cut-off frequency to 625 Hz increased the accuracy on LE speech by 13.4 %. The performance decrease on neutral speech was almost linear with the increasing cut-off frequency (1.8 % for 625 Hz).

9.3.3 Filter Bank Resolution

An ID curve for the FB starting at the optimized cut-off frequency of 625 Hz was analyzed. To obtain a smoothed ID estimate, the number of FB bands was lowered to 12, see Fig. 9.4. The solid line 'comb' in the upper part of the figure represents cut-off frequencies of the FB filters. It can be seen that a major part of the linguistic information is concentrated in the low FB bands, which corresponds

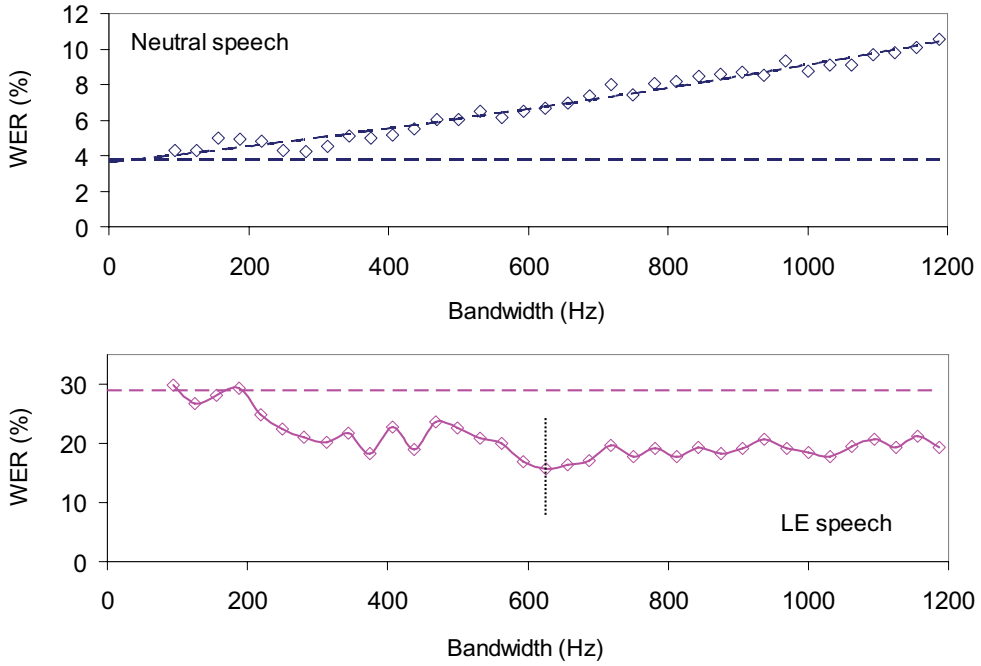

Figure 9.3: Searching for optimal low cut-off frequency in 19-band FB.

Set		Devel set	
		Neutral	LE
WER (%)	LFCC, full band	4.8 (4.1–5.5)	29.0 (27.5–30.5)
	LFCC, ≥ 625 Hz	6.6 (5.8–7.4)	15.6 (14.4–16.8)

Table 9.2: Performance of cepstra derived from a bank of linearly spaced rectangular filters (LFCC): (1) 20 filters, 0–4000 Hz, (2) 19 filters, 625–4000 Hz. Mean values followed by 95% confidence intervals in parentheses.

well with the dominant importance of the first two formants for the speech recognition observed in (Bou-Ghazale and Hansen, 2000).

Subsequently, following the concept used in *Expolog*, FB resolution was increased in the area of the ID curve maximum. The first band in the former 12–bands FB was split into two bands and the former two subsequent bands were substituted with three bands, yielding a 14–bands FB, see dashed line in Fig. 9.4. As a result, surprisingly, the score dropped from 17.2 % to 26.9 % WER for LE. Based on this experiment, it can be concluded that increasing FB resolution in the bands that are most important for speech recognition does not automatically improve the system performance.

9.3.4 Optimizing Frequency Band End-Points

The experiment in changing the FB resolution at low frequencies demonstrated that it is not possible to design an LE-robust FB by just modifying the distribution of bands in the FB according to the ID curve. At the same time, it has been shown that redistributing the filter bands may significantly impact the front-end performance. This observation motivated the development of a FB repartitioning algorithm which will be presented in this section.

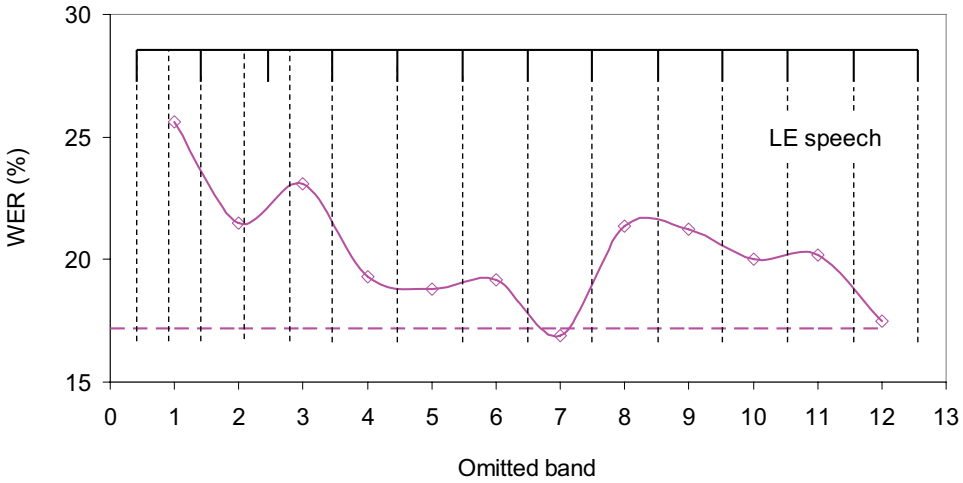

Figure 9.4: Increasing FB resolution in region dominant for speech recognition.

There are many degrees of freedom in the process of optimizing the bandwidths of the FB filters. Even switching order of the filter manipulation steps during the FB design may lead to various semi-optimal FB configurations. E.g., it is likely that modifying parameters of all FB filters at each design iteration will yield different 'optimal' FB than when gradually optimizing parameters of each single filter, keeping the remaining filters intact. Although not promising to yield an optimal FB setup, the latter approach was chosen for the following experiment for further improving the FB performance in the LE-robust feature extraction front-end.

In the proposed repartitioning algorithm, the idea is to search for an optimal bandwidth of each filter while leaving the rest of FB intact as much as possible. In the initial step, an end-point of the first filter was iteratively varied around its original location and a new position yielding minimal WER was searched. For each end-point position, the remaining higher FB bands were resized, each of the filters by the same ratio. Once the optimum for the band was reached (displayed by the minimum in the WER curve), the end-point was fixed. In the subsequent steps, successive bands were processed the same way, keeping the precedent optimized end-points intact.

In order to reduce the computational costs when examining efficiency of the repartitioning algorithm, the number of FB bands was limited to 6. The initial 6–bands FB is depicted in Fig. 9.5 by dotted line, dashed line refers to the initial FB performance. Solid lines represent WERs reached by varying the corresponding frequency band end-points. In the cases when the dotted line passed through the minimum of the WER curve, the initial end-point already occurred in the optimal position and was not further altered.

Decreasing the number of FB bands considerably affected baseline performance, see Table 9.3, nevertheless, the repartitioned FB improved on the baseline 6-bands FB by 2.3 % for LE speech, see Table 9.4.

9.3.5 Evaluation

Showing its efficiency on the 6-bands FB, the repartitioning algorithm was subsequently applied also to the 19-bands FB, reducing WER on LE *devel* set from 15.6 % to 14.8 %. The overall performance of MFCC, PLP, MR-RASTA, and MFCC-based features employing filter banks obtained from the data-driven design was evaluated on the *open* set, see Fig. 9.4. For the baseline features, the results on the *open* set are consistent with the initial *devel* set experiment, see Sec. 9.2.3. MR-RASTA features outperformed both MFCC and PLP on Lombard speech while slightly reduced accuracy on neutral

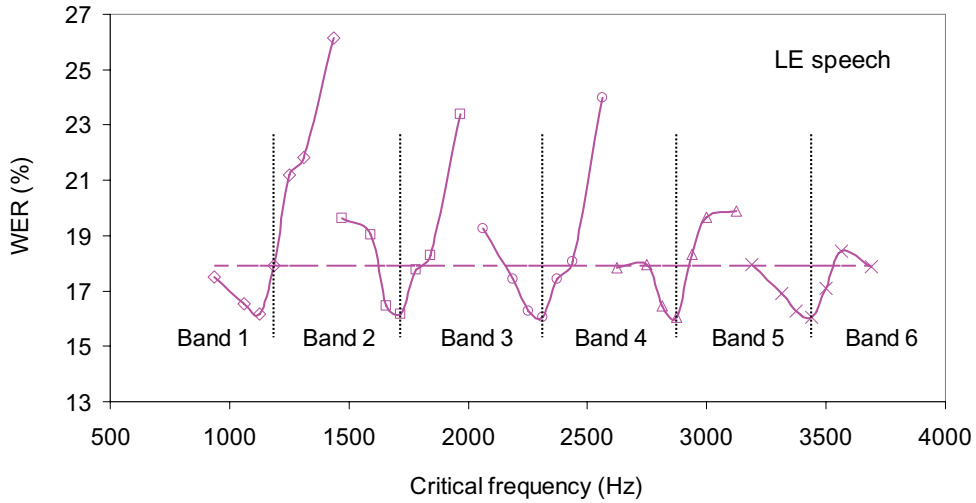

Figure 9.5: Search of optimal band partitioning for 6-band FB. For each band sequentially, end-point yielding best performance is found, preserving distribution of preceding bands.

Set		Devel set	
		Neutral	LE
WER (%)	LFCC, 19 bands	6.6 (5.8–7.4)	15.6 (14.4–16.8)
	LFCC, 12 bands	8.0 (7.1–8.9)	17.2 (15.9–18.5)
	LFCC, 6 bands	9.6 (8.6–10.6)	17.9 (16.6–19.2)

Table 9.3: Performance of cepstra derived from a bank of linearly spaced rectangular filters (LFCC). Mean values followed by 95% confidence intervals in parentheses.

speech (the decrease was not statistically significant). From the newly proposed features, 'LFCC, 20 bands' provided comparable performance to baseline features on neutral speech, while significantly reducing WER of MFCC and PLP on LE speech. On LE speech, the best features were 'LFCC, 19 bands' and 'RFCC, 19 bands'. In the case of 19-bands bank, application of the repartitioning algorithm resulted in a slightly reduced accuracy (statistically insignificant) compared to the baseline 'LFCC, 19 bands'. Since the *open* set was intended to be used also in the further experiments, the *open* test results were not used for picking the best LE-performing FB (which would be 'LFCC, 19 bands'). Rather, the results obtained on the *devel* set were taken as a guideline, suggesting the use of 'RFCC, 19 bands' for the Lombard speech recognition.

Set		Open set	
		Neutral	LE
WER (%)	MFCC	3.7 (2.7–4.7)	68.7 (66.6–70.8)
	PLP	3.4 (2.4–4.4)	61.3 (59.1–63.5)
	MR-RASTA	4.1 (3.0–5.2)	42.1 (39.9–44.3)
	LFCC, 20 bands, full band	3.3 (2.3–4.3)	49.4 (47.1–51.7)
	LFCC, 19 bands, ≥ 625 Hz	6.6 (5.3–7.9)	24.6 (22.7–26.5)
	RFCC, 19 bands, ≥ 625 Hz	6.4 (5.1–7.7)	25.5 (23.5–27.5)
	LFCC, 12 bands, ≥ 625 Hz	7.4 (6.0–8.8)	25.6 (23.6–27.6)
	LFCC, 6 bands, ≥ 625 Hz	9.5 (7.9–11.1)	31.7 (29.6–33.8)
	RFCC, 6 bands, ≥ 625 Hz	8.5 (7.0–10.0)	29.4 (27.3–31.5)

Table 9.4: Evaluation of all systems on open test set: MFCC, PLP, MR-RASTA, cepstra derived from linearly spaced rectangular filters (LFCC) and repartitioned filters (RFCC). Mean values followed by 95% confidence intervals in parentheses.

9.4 Derived Front-Ends

Presence of noise deteriorates the quality of the estimation of speech features in the ASR front-end. In particular, linear predictive coding (LPC) is known to provide excellent resolution properties in spectral modeling, but the presence of noise may considerably alter the spectral estimate, (Kay, 1979). Moreover, LPC seems to be less effective in modeling consonants than DCT, which sometimes results in a reduced ASR performance of LPC-based features compared to the DCT ones, (Davis and Mermelstein, 1980). On the other hand, in (Bou-Ghazale and Hansen, 2000), features extracted from LPC power spectrum provided superior accuracy on clean and noisy stressed speech. According to the authors, the outstanding performance of LPC features resulted from the spectral smoothing, which was able to suppress fine spectral variations caused by the excitation changes when switching between stressed speech styles.

In the following sections, the performance of cepstral features employing various FBs and DCT or LPC-based extraction stages is compared in the CLSD'05 neutral and Lombard digit recognition tasks[7]. The former set of baseline features established in Sec. 9.2 was reduced for MR–RASTA features[8] and extended for Expolog features, see Sec. 9.3, as well as for the following modifications of standard and newly proposed features:

- MFCC–LPC, PLP–DCT: Altered cepstral extraction, DCT in MFCC is replaced by LPC, LPC in PLP is replaced by DCT, respectively,
- 20Bands–LPC: Derived from PLP, former FB is replaced by 20 rectangular filters spread over 0–4 kHz without overlap,

[7]The front-ends comprising modified FB were published in (Bořil *et al.*, 2006c).

[8]Each of the multiple experiments would have required the assistance of the author of the MR–RASTA front-end implementation.

- Big1–LPC: Derived from 20Bands–LPC by merging the first three bands together. This FB modification was motivated by the expectation that decreasing FB resolution in the area that seems to be 'disturbing' for the female LE speech recognition (0—600 Hz), see Sec. 9.3.2, may suppress its negative impact on the LE task, while preserving certain amount of relevant information for the neutral speech recognition,
- RFCC–DCT: Similar to 'RFCC, 19 bands', i.e., FB in MFCC is replaced by the repartitioned 19-bands rectangular FB,
- RFCC–LPC: Derived from PLP, former FB is replaced by the repartitioned 19-bands rectangular FB.

9.4.1 Training Models: One-Step and Progressive Mixture Splitting

Two approaches to the acoustic model training were compared – *one-step mixture splitting* and *progressive mixture splitting*, see Sec. 4.7.1. In *one-step splitting*, after the 9^{th} retraining period, each single Gaussian modeling one element of the feature vector was split to 32 Gaussian mixtures. The acoustic model was then retrained in 60 periods.

In *progressive mixture splitting*, the number of mixtures was doubled repeatedly after several iterations, first after the 9^{th} period, then after every 5 retraining periods until the amount of 32 mixtures was reached. After that, the number of mixtures was kept constant up to a total of 70 retraining periods.

The same *train* and neutral *devel* sets as in Sec. 9.1 were used for the digit recognizer training and evaluation, respectively. For selected features, model convergence in dependency on the number of retraining periods is shown in Fig. 9.6. It can be seen that, with an exception of MFCC, *progressive splitting* provides faster model convergence and better resulting accuracy compared to *one-step splitting*. In the case of MFCC, the model convergence and accuracy reaches similar values for both approaches after approximately 30 retraining periods. In the majority of the front-ends, the acoustic models were sufficiently trained after passing through 50 retraining iterations. In the experiments presented in the following sections, models from the 56^{th} retraining period were used for speech recognition.

9.4.2 Performance on Neutral, LE, and Converted Speech

In this section, recognizers incorporating features from Sec. 9.4 are compared in the neutral and clean LE digit recognition task. In the LE task, the efficiency of the voice conversion-based feature normalization, CLE and $CLEF_0$, see Sec. 8.1, is also evaluated. Only eight-digit sequence utterances had been chosen to be processed by the voice conversion system, hence, a reduced number of items from the *open* set established in Sec. 9.1 was used for the actual tests[9].

Performance of the systems is shown in Table 9.5. The columns denoted CLE and $CLEF_0$ represent setups including voice conversion. The results are also depicted in Fig. 9.7. The following can be observed:

- MFCC reached the poorest results on the LE set compared to the other front-ends. Performance on the neutral set is comparable to Expolog,
- MFCC–LPC: Replacing DCT by LPC in MFCC dramatically improved the front-end's performance on LE speech. On the neutral set, a slight, yet not statistically significant increase of accuracy was reached. MFCC-LPC displayed a performance comparable to PLP on both sets,

[9]In addition, one of the neutral sessions contained reduced number of digit sequence utterances, resulting in a lower number of items in the neutral *open* set compared to the LE and converted sets.

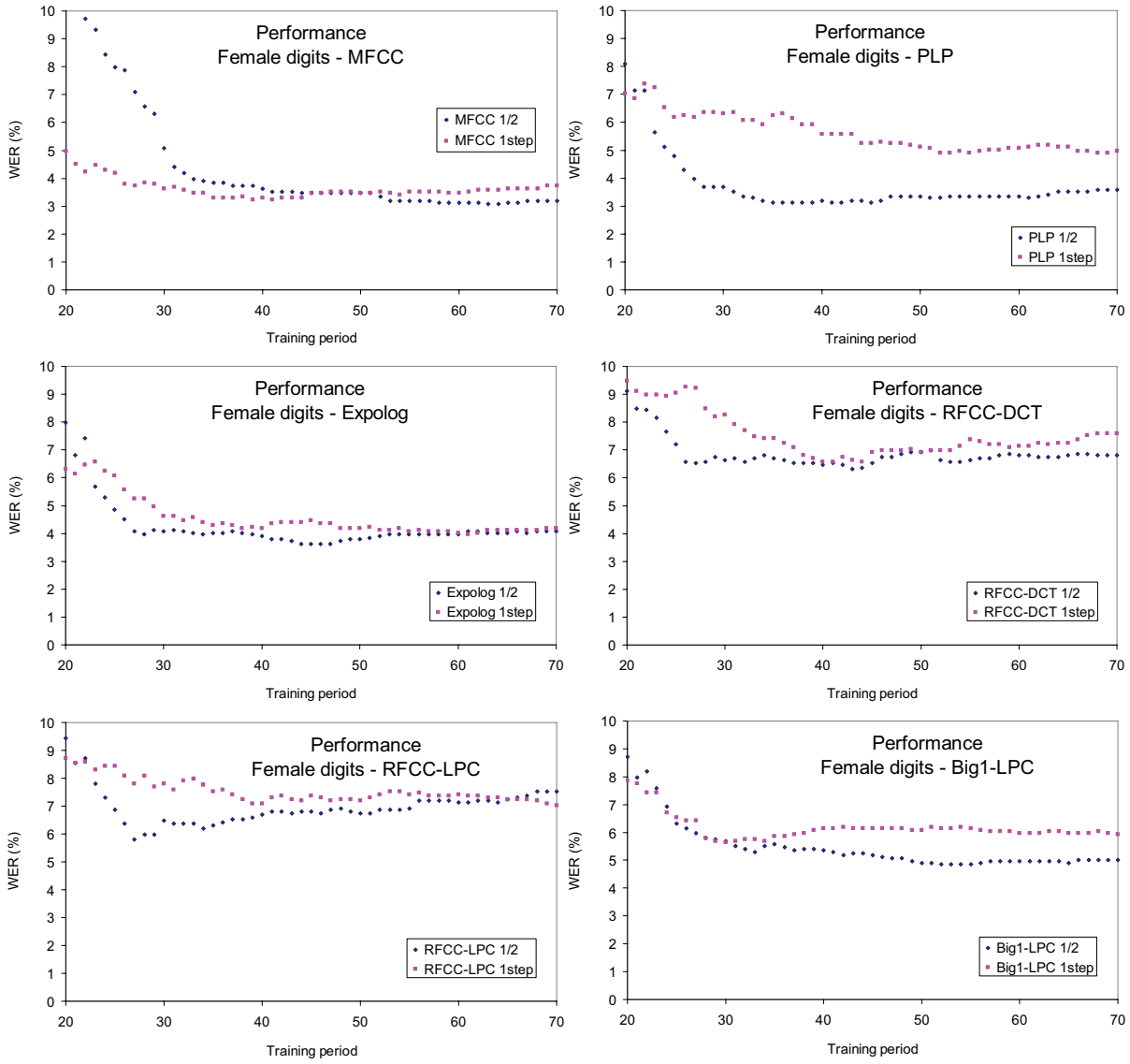

Figure 9.6: Comparing efficiency of monophone models training approaches – one-step (1step) and progressive mixture splitting (1/2). Tests performed on neutral digits development set.

- PLP–DCT: Replacing LPC by DCT in PLP resulted in a slight performance drop on LE speech,
- Expolog features considerably outperformed MFCC, PLP, as well as their LPC/DCT variations. On the neutral set, the performance was similar to MFCC,
- 20Bands–LPC displayed comparable accuracy to PLP on neutral speech, while outperforming MFCC, PLP, and their variations on LE speech,
- Big1–LPC outperformed all aforementioned front-ends on the LE set, at the expense of efficiency on the neutral set. As expected, reducing the FB resolution in the interval 0–600 Hz helped to suppress an impact of frequency components corrupting LE speech recognition. On the other hand, part of the information important for the neutral speech recognition was lost this way,

Set		Neutral	LE	CLE	$CLEF_0$
# Digits		768	1024	1024	1024
WER (%)	MFCC	3.7 (2.3–5.0)	71.3 (68.5–74.1)	30.9 (28.0–33.7)	58.6 (55.6–61.6)
	MFCC-LPC	2.9 (1.7–4.0)	46.7 (43.6–49.7)	26.2 (23.5–28.9)	48.8 (45.8–51.9)
	PLP	2.9 (1.7–4.0)	47.4 (44.3–50.4)	25.2 (22.5–27.9)	49.0 (46.0–52.1)
	PLP-DCT	2.5 (1.4–3.6)	51.2 (48.1–54.2)	23.4 (20.8–26.0)	52.5 (49.5–55.6)
	Expolog	3.9 (2.5–5.3)	35.8 (32.8–38.7)	26.7 (24.0–29.4)	37.8 (34.8–40.8)
	20Bands-LPC	3.0 (1.8–4.2)	42.1 (39.1–45.1)	19.9 (17.5–22.4)	42.7 (39.7–45.7)
	Big1-LPC	4.7 (3.2–6.2)	27.2 (24.4–29.9)	20.4 (17.9–22.9)	25.3 (22.6–28.0)
	RFCC-DCT	5.1 (3.5–6.6)	26.1 (23.4–28.8)	22.6 (20.0–25.1)	22.6 (20.0–25.1)
	RFCC-LPC	4.6 (3.1–6.0)	23.0 (20.4–25.5)	23.1 (20.6–25.7)	23.7 (21.1–26.3)

Table 9.5: Comparing front-ends: Efficiency of voice conversion-based LE normalization. Mean values followed by 95% confidence intervals in parentheses.

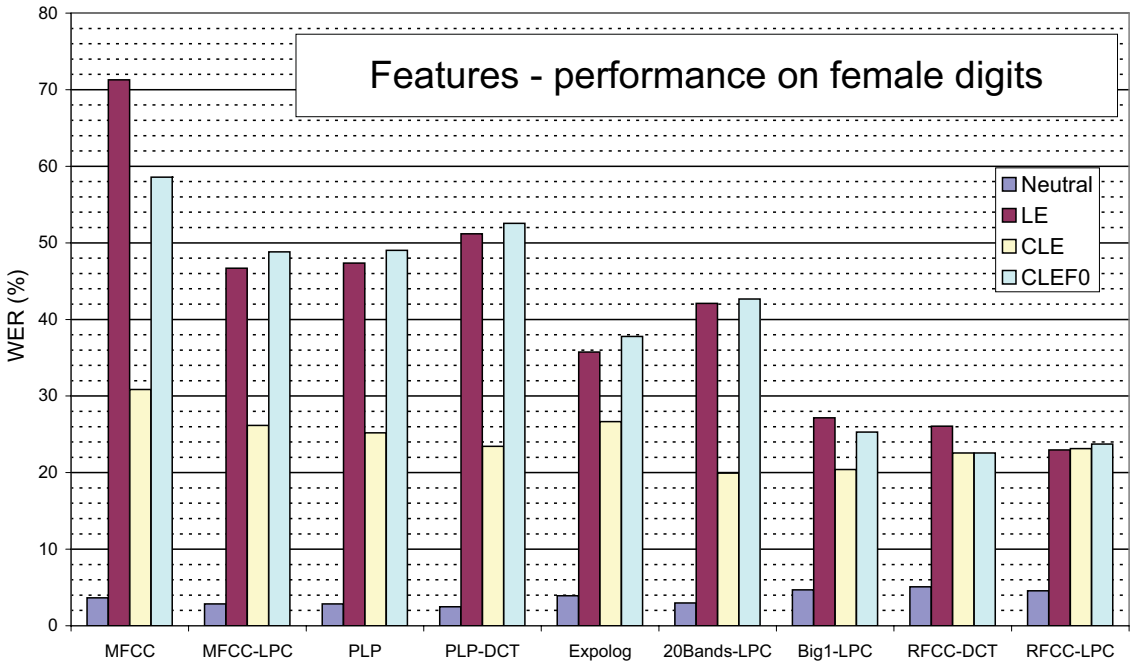

Figure 9.7: Comparing front-ends: Efficiency of voice conversion-based LE normalization.

- RFCC–DCT/LPC: MFCC and PLP-based features employing repartitioned 19-bands FBs displayed superior performance on the LE set. RFCC–LPC reached the best LE performance compared to other front-ends while its accuracy on neutral speech was similar to Big1–LPC.

It can be seen that in all cases the LPC-based cepstral coefficients provided better robustness to LE compared to the DCT-based coefficients. In the case of neutral speech, either replacing DCT by LPC (MFCC–LPC, RFCC–LPC) or LPC by DCT (PLP–DCT) slightly improved performance of the system, but the changes were not statistically significant.

PLP features displayed less sensitivity to the DCT/LPC alternation compared to MFCC. The same was true for the PLP-derived front-end RFCC.

Big1–LPC was designed to preserve a certain portion of the information relevant for neutral speech recognition, which was discarded in RFCC by omitting the frequency band 0–600 Hz. However, Big1–LPC and RFCC–LPC features displayed similar performance on the neutral set. Based on the experiments in Sec. 9.3.2, it can be assumed that increasing the FB resolution at the low frequencies would improve Big1–LPC performance on neutral speech at the expense of LE speech recognition.

20Bands–LPC seems to be a promising replacement for MFCC and LPC features, providing similar performance on neutral speech and significantly better performance on Lombard speech.

Applying voice conversion to both formants and F_0 improved the system performance for the majority of front-ends, confirming observations made in Sec. 8.2. The best performance was reached when combining CLE and 20Bands–LPC or Big1–LPC. In the case of features focusing on the robustness to LE (Expolog, Big1–LPC, RFCC), the relative improvement of LE recognition was less significant than in the standard features. Voice conversion applied just to F_0, $CLEF_0$, did not display any consistent and significant improvements across the front-ends. This is presumably due to the fact that both the F_0 and formant components of the amplitude spectra shift significantly under LE, see Sec. 6.4 and Sec. 6.5. Normalization of F_0 may help to reduce the F_0 interference with the location of F_1 typical occurrence, however, it does not address mismatch of the Lombard formant structure and neutral-trained acoustic model.

9.4.3 Performance as a Function of Fundamental Frequency

The CLSD'05 database comprises utterances recorded in the constant level of the simulated background noise. However, partly due to the imperfection of the recording setup where the volume of the speech feedback sent to the headphones was adjusted subjectively according to the demands of speakers, partly due to the fact that individuals tend to react to the similar noise differently, it can be assumed that CLSD'05 captures LE of various levels[10].

Past studies reported that average vocal intensity in humans displayed a linear dependency on the level of the noise background (although different slopes were found for the 'just reading text' and the 'intelligible communication over noise' setups), (Webster and Klumpp, 1962), (Lane *et al.*, 1961) (see Sec. 3.2.1 for details), and that average F_0 in semitones linearly increased with vocal intensity, see Sec. 3.2.2. When dealing with an utterance set containing a mixture of neutral and LE speech, it can be expected, making a crude approximation, that the higher is the average F_0 of the particular utterance, the higher is the probability that the utterance comprises Lombard speech rather than neutral speech.

In this section, the dependency of the recognition performance on the average utterance's F_0 is evaluated for a selected subset of the previously compared front-ends, assuming that with increasing F_0 the characteristic may reflect, to a certain extent, the systems' resistance to the increasing level of LE.

The female neutral and LE digit sets used in the previous section were merged, yielding a set of utterances with F_0 covering an interval of approximately 100–500 Hz, see Fig. 9.8. To obtain a reasonable amount of data for each point of the measured $WER(F_0)$ characteristic, utterances with F_0 falling within a sliding window of the bandwidth 100 Hz were assigned to the test set F_c, where F_c denotes the central frequency of the window. The window was shifted with a step of 5 Hz, starting

[10]So far, the term *level of LE* has not an exact, generally accepted definition. In this thesis, the *level of LE* is used in the relation to the ratio of shift of a certain speech feature or of a set of features from their neutral positions due to LE.

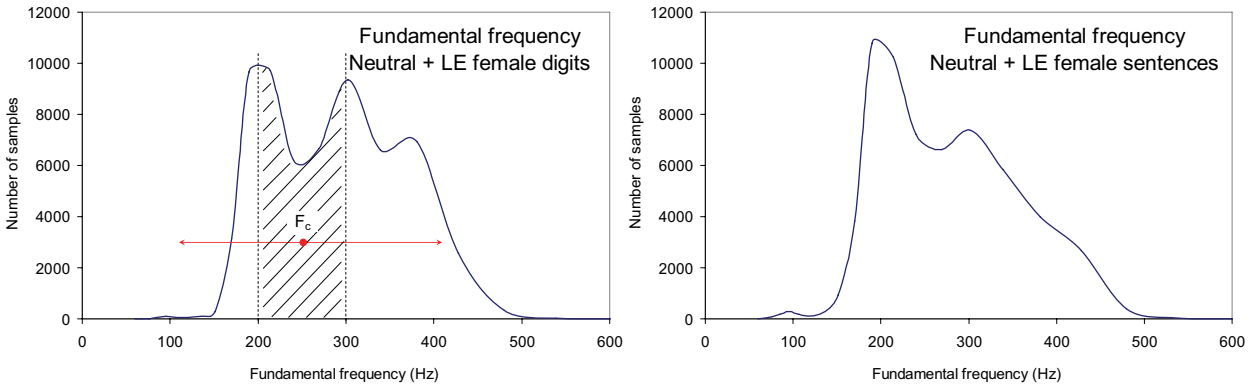

Figure 9.8: F_0 distributions of female open sets – merged neutral and LE utterances.

from F_c = 150 Hz. At each step, a recognition was performed for the actual F_c set to sample the WER(F_c) curve. Results of the experiment are shown in Fig. 9.9. For F_c starting at 250 Hz (F_0

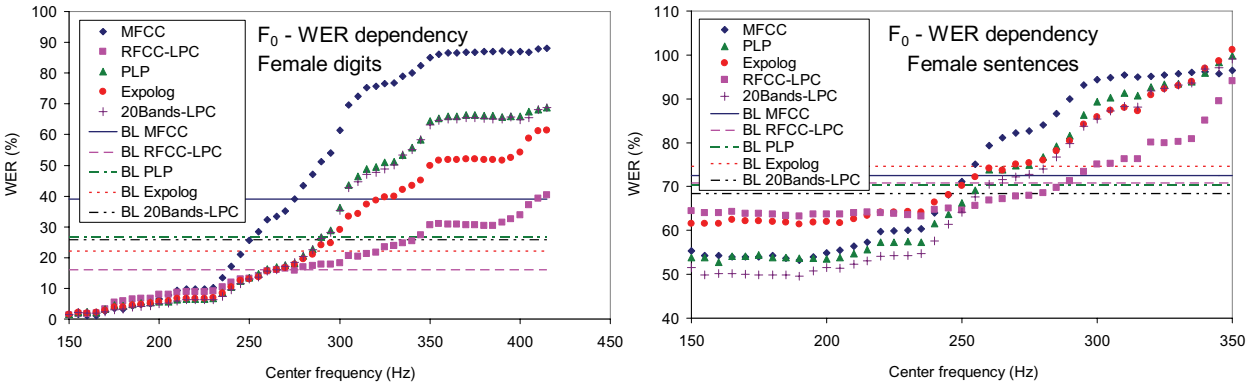

Figure 9.9: WER(F_c) dependency. BL – baseline WER on the whole merged neutral + LE set.

= 200–300 Hz), RFCC–LPC outperforms the other features. This observation is consistent with the RFCC–LPC superior robustness to LE reached in the baseline test, see Fig. 9.7. When decreasing F_c below 250 Hz, PLP, 20Bands–LPC, and MFCC display the best performance.

The same recognizer as in the digit task was employed in the LVCSR experiment, only its vocabulary was extend to 1095 words[11]. Phonetically rich sentences uttered by the digit open set speakers, see Sec. 9.1, formed the LVCSR open set. Distribution of the merged female neutral and LE sentences is shown in Fig. 9.8. Since the framework comprised no language model (word transition probabilities were set uniform), the baseline performance of the system was quite poor in this case. However, the performance of the front-ends being exposed to the sentences with increasing average F_0 display similar trends as in digits, see Fig. 9.9. For F_c approximately 250 Hz and higher, i.e., for the set of utterances of average F_0 from 200 to 300 Hz[12], both on digits and sentences RFCC–LPC starts to outperform the other features, while 20Bands–LPC followed by PLP and MFCC reach the best results for F_c bellow 250 Hz. This observation is consistent with the results of the baseline test, see Sec. 9.4.2. The present experiment indicates that RFCC–LPC can be expected to provide a superior performance

[11] The TUL LVCSR recorgnizer described in Sec. 4.7.2 was not available for this experiment.

[12] Average F_0 of CLSD'05 female utterances lies around 200 Hz for neutral and around 300 Hz for LE speech, see Table 6.2, 8.2.

both in digit and LVCSR tasks when exposed to LE of various levels. Moreover, the results suggest building a system that would employ different front-ends for neutral and Lombard speech recognition. This approach will be further explored in Chap. 11.

9.4.4 Performance in Noisy Conditions

In this section, feature robustness to various types and levels of additive noise is tested. One noise sample from the Car2E database (Pollák *et al.*, 1999) and five noises from the Aurora 2 database (Hirsch and Pearce, 2000) were mixed with the clean open set digit utterances from Sec. 9.4.2 at SNR from -5 to 20 dB with a step of 5 dB.

Car2E noises were recorded in the cabin of a moving car and labeled according to their 'stationarity'. The noise sample picked for the actual experiment took place in the set of noises used for the CLSD'05 Lombard scenario recordings and belongs to the category of the most stationary signals. From the Aurora 2 database, the following noises were used: crowd of people (babble), restaurant, street, airport, and train station.

Performance of the front-ends from Sec. 9.4 was tested in the noisy digits task. Moreover, efficiency of noise subtraction (NS) based on the full-wave rectified spectral subtraction and Burg's cepstral VAD, (Fousek, 2007)[13], as implemented in CTUCopy open source tool, was evaluated. The SNR–WER dependencies for three best performing and two worst performing front-ends in car and babble noises are shown in Fig. 9.10. Since the SNR–WER curves were often interlaced, the features outperforming the others in most SNRs or displaying consistent superior performance in lower SNRs were chosen as the best features. In the figure, 'INF' denotes a condition, where no noise was added to the clean speech signal[14]. The left part of the figure shows performances without NS and the right part performances with NS included. Three best performing features for neutral and Lombard speech without using NS are listed for each of the scenarios in Table 9.6.

NS employed in the presented experiments relies on the ability of VAD to detect non-speech parts of the signal. In general, increasing level of noise in the signal reduces VAD's accuracy. As a consequence of the VAD's errors, noise estimates subtracted from the signal may contain also portions of speech spectra, causing NS to degrade the speech signal. From a certain level of noise, the ASR performance reduction introduced by failing NS may be stronger than the performance reduction caused by the presence of noise in the former signal. Hence, noise levels up to that where NS improved the recognition performance were searched (see the fourth column of Table 9.6 denoted 'NSeff').

Set	Neutral	LE	NSeff (dB)
Airport	MFCC, 20Bands–LPC, PLP	Big1–LPC, RFCC–LPC, Expolog	None
Babble	MFCC, MFCC–LPC, PLP–DCT	RFCC–LPC, Expolog (Big1–LPC), RFCC–DCT	10
Car2E	Expolog, 20Bands–LPC, Big1–LPC	RFCC–LPC, Big1–LPC, Expolog	-5
Restaurant	MFCC, 20Bands–LPC, MFCC–LPC	RFCC–LPC, Big1–LPC, RFCC–DCT	-5
Street	20Bands–LPC, MFCC, Expolog	RFCC–LPC, Big1–LPC, 20Bands–LPC	0
Train station	20Bands–LPC, MFCC, Expolog	RFCC–LPC, Big1–LPC, 20Bands–LPC	-5

Table 9.6: Features performing best on neutral noisy or LE noisy speech.

Apparently, the choice of optimal features depends heavily on the type of background noise. Presented experiments show that both DCT or LPC-based cepstral coefficients may be an effective choice for the recognition of neutral speech, depending on the actual noise characteristics. For LE speech,

[13]VAD based spectral subtraction is briefly discussed in Sec. 4.1.2, see the second footnote.
[14]Mean SNR of the CLSD'05 clean neutral speech recordings is approximately 28 dB, see Sec. 6.3.

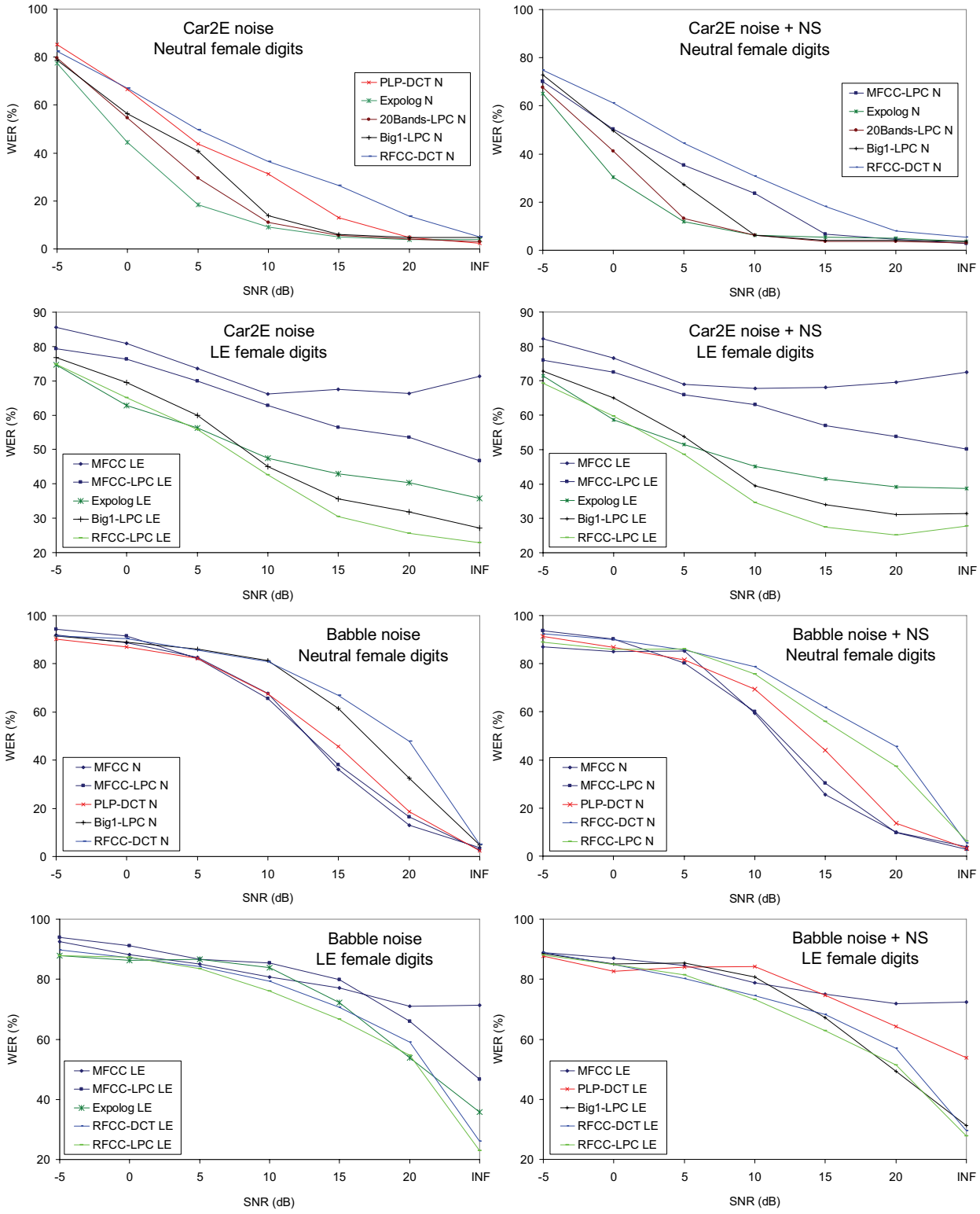

Figure 9.10: Three best and two worst-performing features in Car2e and babble noise.

LPC-based features displayed superior performance in all types of noise background. Here, the best performance was reached by RFCC–LPC for almost all cases.

9.5 Conclusions

In this chapter, new features for robust Lombard speech recognition were proposed, employing FBs obtained in the data-driven design. The goal of the FB design was to emphasize spectral components carrying the dominant portion of linguistic information in Lombard speech and to suppress components that disturb the recognition.

The proposed features were compared in digit tasks to MFCC and PLP, to their modifications MFCC–LPC and PLP–DCT, and to the state-of-the-art Expolog and MR–RASTA features. The novel RFCC–LPC displayed consistently superior performance on clean Lombard speech, noisy Lombard speech, and speech with F_0 exceeding the mean F_0 of female neutral speech as observed in CLSD'05. 20Bands–LPC considerably outperformed MFCC and PLP on Lombard speech while preserving comparable performance on neutral speech. Since 20Bands–LPC employs a filter bank of 20 rectangular filters distributed equidistantly on the linear frequency axis, this result suggests that feature extraction filter banks derived from auditory models do not necessarily represent the optimal choice for ASR.

Effectiveness of voice conversion-based speech normalization when combined with the standard and newly proposed features was also evaluated in the digit task. Voice conversion substantially improved performance of MFCC and PLP and their variations. In the case of the newly proposed features, with the exception of 20Bands–LPC, voice conversion was less effective (Big1–LPC, RFCC-DCT) or not effective at all (RFCC–LPC). Note that voice conversion, presumably due to its limited accuracy, was not successful in improving performance of the LVCSR system, see Sec. 8.3.

In the case of neutral speech, cepstral coefficients derived either from LPC or DCT displayed good performance, depending on the type of front-end. On Lombard speech, LPC-based cepstra systematically outperformed DCT cepstra in all front-end setups. This result is consistent with the observation made previously by (Bou-Ghazale and Hansen, 2000), where LPC-based spectral smoothing suppressed fine spectral variations caused by LE-induced excitation changes.

Chapter 10

Frequency Warping

The location of vocal tract resonances, formants, is inversely proportional to the vocal tract length (VTL), (Lee and Rose, 1996). During speech production, the resonance frequencies and bandwidths are modified by articulators to produce distinct speech sounds. VTL differs across individuals, genders, and age groups, ranging approximately from 18 cm for males to 13 cm for females, (Lee and Rose, 1996), and down to 7 cm for new-born babies, (Vorperian *et al.*, 2005). VTL differences introduce considerable inter-speaker variability of formant structure which may cause a significant degradation of ASR performance when switching from speaker-dependent to speaker-independent acoustic models. In the past decade, vocal tract length normalization (VTLN) has become a popular means to address this source of variability. In VTLN, short-time spectra of speech are normalized by applying speaker-dependent frequency warping. The warping is usually conducted using a scalar factor α, where the warped frequency scale F_W is derived from the original scale F as $F_W = F/\alpha$.

As previously shown in Sec. 6.5, LE introduces considerable shifts of formants from the locations seen in neutral speech. The extent and consistency of the formant structure shifts (systematic shifts of F_1 and F_2 to higher frequencies) suggest that VTLN techniques might be efficient in normalizing LE formants towards neutral. In this chapter, VTLN-based Lombard speech normalization is incorporated in the feature extraction front-end and tested in the ASR task. Two novel methods of the formant structure normalization derived from the maximum likelihood approach and parametric approach to VTLN are proposed in the following sections.

In the standard *maximum likelihood approach* (ML–VTLN), warping factor α maximizing the likelihood of the normalized acoustic observation $\boldsymbol{O}^{\alpha}$ is searched given a transcription W and hidden Markov model $\boldsymbol{\Theta}$, (Lee and Rose, 1996), (Pye and Woodland, 1997), (Garau *et al.*, 2005):

$$\hat{\alpha} = \arg\max_{\alpha} \left[\Pr\left(\boldsymbol{O}^{\alpha} | \, \boldsymbol{W}, \boldsymbol{\Theta} \right) \right]. \tag{10.1}$$

In the *parametric approach*, the warping factor is derived from the estimated formant positions, (Eide and Gish, 1996). VTLN is usually applied both during training, yielding speaker-independent, frequency-normalized acoustic models, and during recognition, normalizing spectra of the test utterances towards the normalized models.

10.1 ML–Based VTLN: Background

In the standard ML–VTLN, (Lee and Rose, 1996), (Pye and Woodland, 1997), the optimal speaker-dependent α is estimated in the iterative procedure where the speech signal is warped consecutively by a set of candidate α's. When applying VTLN during the HMM training (*VTLN training*), multiple

forced alignment is conducted on the training utterances using α's from the candidate set[1]. For each utterance, the α maximizing likelihood of the forced alignment is picked as an estimate of the optimal utterance warping. Optimal α for the actual speaker is estimated to maximize overall alignment likelihood over all speaker's training utterances. Once the procedure is finished for all speakers, the training set is warped by the speaker-dependent α's and used for re-training the acoustic models.

In the recognition stage, unwarped test utterances are first decoded using the normalized HMMs, yielding estimates of the test transcriptions. Using these transcription estimates, utterance-dependent α's are then estimated similarly as in the training stage. Finally, the utterances are warped using the utterance-dependent α's and decoded (*VTLN recognition*).

Frequency warping can be realized either in the time domain by resampling the signal waveform, (Pye and Woodland, 1997), (Sündermann *et al.*, 2005c), or, more commonly, in the frequency domain by stretching or shrinking the filter bank cut-offs in the feature extraction front-end, (Lee and Rose, 1996), (Pye and Woodland, 1997).

10.2 ML–Based VTLN: A Novel Approach

In the following experiments, the filter bank-based warping was used. Efficiency of two setups of *VTLN training* was evaluated. In the first setup, standard *VTLN training* as described above was carried out, including speaker-dependent warping of the training utterances. In the second setup, α's estimated on the utterance level were directly used for warping the corresponding utterances instead of using global, speaker-dependent α's. It is assumed that the first setup will provide a more reliable estimate of the speaker's 'mean' VTL as being derived from multiple samples, while the second setup may better capture temporal deviations of the speaker's VTL due to the changes in the configuration of articulators. The *VTLN training* procedure comprised the following steps:

- Training HMMs in 36 iterations, progressive mixture splitting, no VTLN,
- Multiple forced alignment for $\alpha \in \{0.8, 0.85, 0.9, \ldots, 1.2\}$, picking α's maximizing likelihood of the forced alignment,
- Utterances warped by utterance-dependent or speaker-dependent α's, feature extraction,
- Retraining HMMs with warped utterances, 3 iterations.
- Repeating step 3,
- Retraining HMMs in 7 iterations $\rightarrow$ normalized HMMs.

Two versions of *VTLN recognition* were proposed and tested in the presented experiments – *utterance-dependent warping approach* and *speaker-dependent warping approach*. Since LE severely corrupts accuracy of speech recognition, see Sec. 8.2, 8.3, the transcriptions estimated by decoding the unwarped utterances considerably differ from the real transcriptions and might be rather confusing than helpful in finding the optimal warping factors. For this reason, the standard *VTLN recognition* procedure was modified as follows:

- Test utterance i was consecutively warped by $\alpha \in \{0.8, 0.85, 0.9, \ldots, 1.2\}$ and for each warping decoded using the normalized HMMs, yielding a set of ML transcription estimates $\left\{\hat{W}_{i,\alpha_1}, \hat{W}_{i,\alpha_2}, \ldots, \hat{W}_{i,\alpha_N}\right\}$ and a set of transcription likelihoods

$$\{P_{i,\alpha_1}, P_{i,\alpha_2}, \ldots, P_{i,\alpha_N}\} = \left\{P\left(\hat{W}_{i,\alpha_1}\right), P\left(\hat{W}_{i,\alpha_2}\right), \ldots, P\left(\hat{W}_{i,\alpha_N}\right)\right\}, \tag{10.2}$$

[1] In the tasks focusing on adult speakers, the candidates for α can be typically found in the interval of 0.8–1.2, (Faria and Gelbart, 2005).

where $\alpha_1 = 0.8, \alpha_2 = 0.85, \ldots, \alpha_N = 1.2$,

- For each utterance, α providing the highest likelihood transcription was picked as the optimal warping estimate:

$$\hat{\alpha}_i = \arg\max_{\alpha} \{P_{i,\alpha}\}, \tag{10.3}$$

- Subsequently, in the *utterance-dependent warping approach*, $\hat{W}_{i,\hat{\alpha}_i}$ was picked as the resulting transcription estimate,

- In the *speaker-dependent warping approach*, the optimal speaker-dependent α was estimated as the median value of the speaker's optimal utterance warping estimates:

$$\hat{\alpha}_{sp} = \operatorname*{med}_{i \in \mathcal{I}_{\mathcal{SP}}} \{\hat{\alpha}_i\}, \tag{10.4}$$

  where $\mathcal{I}_{\mathcal{SP}}$ is the set of indexes of all utterances from the given speaker. For each utterance, $\hat{W}_{i,\hat{\alpha}_{sp}}$ was then picked as the final speaker-dependent transcription estimate.

Similarly as in the case of *VTLN training*, the introduction of speaker-dependent warping here is motivated by an effort to obtain more reliable estimate of the speaker's VTL compared to the utterance-dependent approach. The *speaker-dependent warping approach* assumes that the speaker identity is known for the actual test utterance and that warping estimates from multiple speaker's utterances are available prior to the actual utterance decoding. Both assumptions were fulfilled in the case of the presented experiments as the assignment speaker–utterance was known and all test data were available at the same time. Considering a real-world application, the first assumption would require inclusion of a speaker identification algorithm into the ASR system. The second assumption could be met by storing $\hat{\alpha}_i$'s from the previously uttered sentences, these values would be then used for finding the actual $\hat{\alpha}_{sp}$ in Eq. (10.4).

All experiments were carried out on both female and male speech. In the case of female experiments, 37 female office sessions from Czech SPEECON were used for neutral/gender-dependent HMM training, and 2560 digits/10 speakers per scenario (neutral, LE) were used for tests. In male experiments, 30 Czech SPEECON sessions were used for gender-dependent training and 1423 neutral digits and 6303 LE digits from 14 speakers were used for tests. Similarly as in Sec. 9.1, all data were down-sampled from 16 kHz to 8 kHz and filtered by G.712 telephone filter.

10.2.1 VTLN Recognition

It can be expected that formant structure normalization provided by *VTLN training* will result in the reduction of cepstral features variation and, hence, in more selective acoustic models compared to those trained without VTLN (unnormalized models). On the other hand, unnormalized models capture greater speech variability and in the *VTLN recognition* stage, it may be easier to find a warping providing a good match of the actual acoustic observations with the 'broader' model characteristics. To evaluate, which kind of acoustic models deals better with neutral and LE speech, a system employing *VTLN recognition* with unnormalized models and a system employing both *VTLN training* and *VTLN recognition* were tested. This section focuses on the first case.

In the speech data sampled by 8 kHz, the area relevant for ASR is restricted to 0–4 kHz. To allow for both spectral shrinking and stretching in the range of 0.8–1.2 without exceeding the upper limit of 4 kHz, the initial filter bank (FB) for unwarped feature extraction was chosen to be spread over 0–3200 Hz. The FB consisted of 20 equidistant rectangular filters without overlap, replacing the standard bark-distributed trapezoidal FB in PLP. This feature extraction setup was chosen due to its superior properties displayed in the previous experiments, see Sec. 9.4.2. Unnormalized HMMs were

trained in 46 iterations using progressive mixture splitting. During the recognition, efficiency of both *utterance-dependent* and *speaker-dependent* VTLN was evaluated.

Distributions of utterance dependent α estimates ($\hat{\alpha}_i$) for female and male sets are shown in Fig. 10.1, 10.2. Values of $\alpha > 1$ correspond to the FB shrinking (endpoints are shifted down to F/α) which results, from the viewpoint of cepstral coefficients extracted from the FB outputs, in the stretching of speech amplitude spectrum, and values of $\alpha < 1$ correspond to the FB stretching which causes spectral shrinking. It can be seen that for both female and male speakers, the maxima of the α

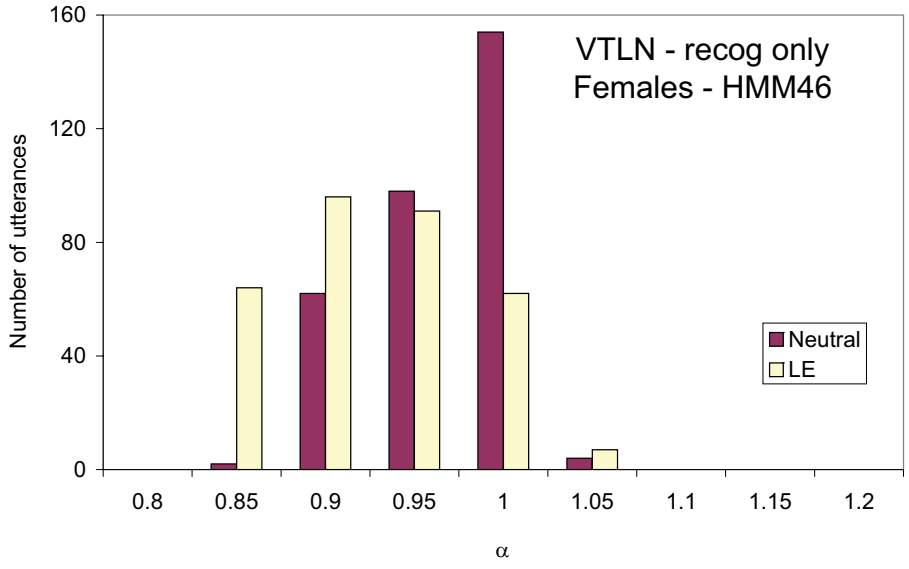

Figure 10.1: Distribution of utterance-dependent α – females, neutral and LE open sets, gender-dependent models.

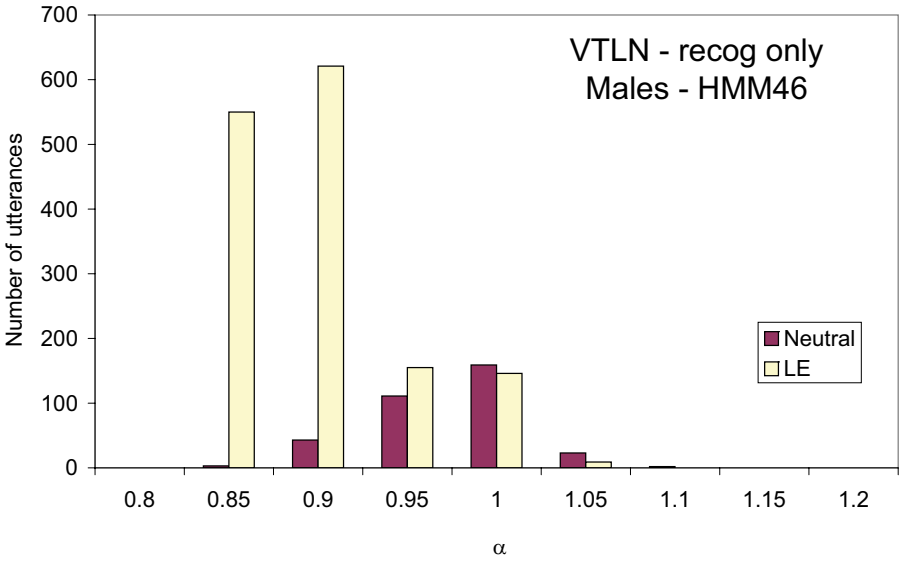

Figure 10.2: Distribution of utterance-dependent α – males, neutral and LE open sets, gender-dependent models.

distributions appear at $\alpha = 1$ for neutral speech and at $\alpha = 0.9$ for LE speech. This suggests that

the spectrum of LE speech needs to be shrunk to better match the neutral acoustic models. This result corresponds with the intuition that formants shifted up the frequency axis due to LE, see Sec. 6.5, should be shifted back down in order to fit the neutral models. Surprisingly, the means of the α distributions for neutral female and male speech reach obviously values $\alpha < 1$, although it might be expected that neutral test speech data would match best the neutral models when being unwarped ($\alpha = 1$). This phenomenon will be further studied in the following sections.

Performance of the *utterance-dependent* and *speaker dependent* VTLN recognition is shown in Table 10.1. The 'baseline' row refers to the recognition where no warping was conducted ($\alpha = 1$). It can be seen that *VTLN recognition* setup provides significant improvement over the baseline system for both female[2] and male Lombard speech. On neutral speech, the improvements are also consistent but do not reach statistical significance. *Utterance-dependent VTLN* and *speaker-dependent VTLN* display similar performance. Hence, *utterance-dependent VTLN* seems to be a preferable choice for *VTLN recognition* due to its lower complexity.

Set		Females		Males	
		Neutral	LE	Neutral	LE
# Digits		2560	2560	1423	6303
WER (%)	Baseline	4.3 (3.5–5.0)	33.6 (31.8–35.5)	2.2 (1.4–2.9)	22.9 (21.8–23.9)
	Utterance-dependent VTLN	3.7 (2.9–4.4)	26.9 (25.2–28.6)	2.0 (1.2–2.7)	20.9 (19.9–21.9)
	Speaker-dependent VTLN	3.8 (3.1–4.5)	27.3 (25.6–29.1)	1.7 (1.0–2.4)	20.5 (19.5–21.5)

Table 10.1: Performance of speaker-dependent and utterance-dependent VTLN recognition, HMM46. Mean values followed by 95% confidence intervals in parentheses.

10.2.2 Joint VTLN Training and VTLN Recognition

In this section, a system conducting both *VTLN training* and *VTLN recognition* is tested. Efficiency of two versions of *VTLN training* employing *speaker-dependent* and *utterance-dependent* warping as defined in the beginning of Sec. 10.2 is compared. In the recognition stage, *speaker-dependent VTLN* and *utterance-dependent VTLN* recognition is carried out.

Distributions of the utterance-dependent α's for the training set as observed during the first and second VTLN retraining stage (corresponding to the model sets HMM36 and HMM39) are shown in Fig. 10.3, 10.4. In the iteration HMM36, the maximum of the α distribution is reached for $\alpha = 1$, while the distribution mean is obviously located at $\alpha < 1$, similarly as in the *VTLN recognition* experiment in the previous section, see Fig. 10.1, 10.2. In the second VTLN retraining iteration (HMM39), the maximum of the α distribution shifts to $\alpha < 1$ for both male and female neutral training utterances. Since the same data sets were used for the model training and VTLN retraining, the requirement of VTLN to systematically warp the data is very surprising. This phenomenon is further studied in the following section.

Results of the joint *VTLN training* and *VTLN recognition* are shown in Table 10.2. The 'baseline' row refers to the task where unwarped test data ($\alpha = 1$) were decoded using the normalized HMM. Compared to the *VTLN recognition* experiment employing unnormalized models, see Table 10.1, the

[2]Note that the baseline performance on the female LE speech reached here is considerably higher than the performance of 'LFCC, 20 bands, full band' on the open set in Table 9.4, Sec. 9.4. This results from the fact that in Sec. 9.4, the open set was more adverse to ASR than the development set, see Table 9.3.2. In the actual VTLN experiments, the open set comprises merged subsets of both development and open set from Sec. 9.4.

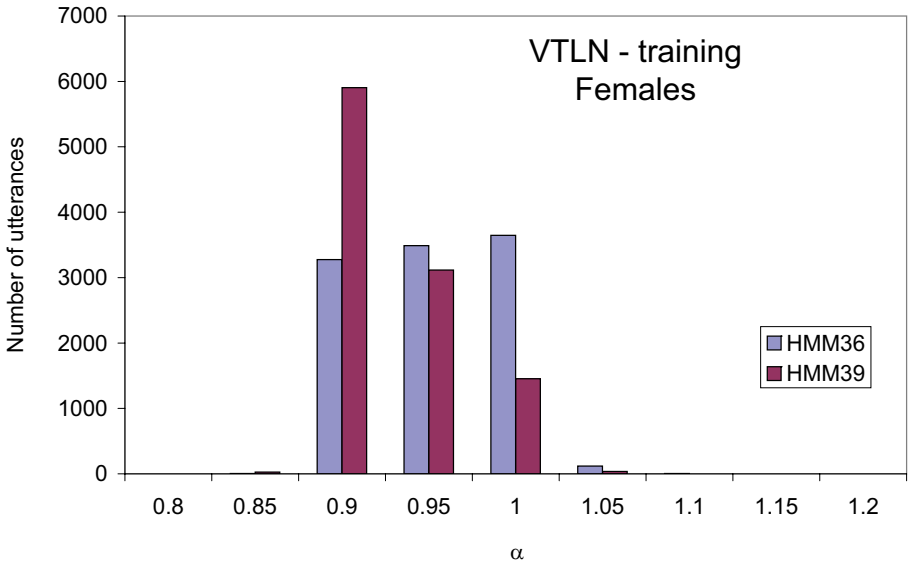

Figure 10.3: Distribution of utterance-dependent α – female neutral train set, retraining iter. 36, 39.

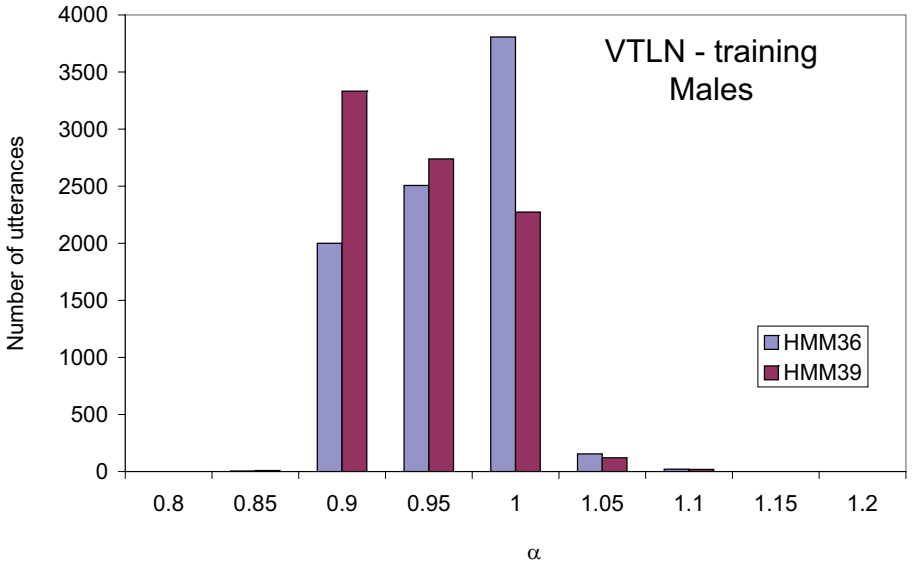

Figure 10.4: Distribution of utterance-dependent α – male neutral train set, retraining iter. 36, 39.

performance of the actual baseline system is considerably worse. This could have been expected since the probability density functions of the unnormalized models were extracted from the unnormalized speech and, hence, better capture its feature distributions compared to the VTLN models trained on the data of reduced variance.

The joint *VTLN training* and *VTLN recognition* approach considerably outperforms the baseline setup. For female neutral and Lombard speech and male neutral speech the joint approach provides similar performance to the *VTLN recognition* using unnormalized models while for the male Lombard speech it brings further, statistically significant improvement of accuracy.

Set		Females		Males	
		Neutral	LE	Neutral	LE
# Digits		2560	2560	1423	6303
WER (%)	Baseline	5.6 (4.7–6.5)	43.8 (41.9–45.8)	2.7 (1.9–3.6)	26.7 (25.6–27.8)
	Utterance-dependent VTLN	3.6 (2.9–4.3)	28.2 (26.4–29.9)	1.8 (1.1–2.4)	16.6 (15.7–17.6)
	Speaker-dependent VTLN	4.0 (3.2–4.7)	27.7 (26.0–29.5)	1.8 (1.1–2.4)	17.4 (16.5–18.3)

Table 10.2: Performance of joint VTLN training and VTLN recognition, same type of VTLN was applied during training and recognition, HMM46. Mean values followed by 95% confidence intervals in parentheses.

10.2.3 VTLN Warping Trends

In Sec. 10.2.1, the neutral open test set was warped during *VTLN recognition* to better match the neutral trained acoustic models. Since the train and test sets were disjunct, the requirement for warping might have been caused by the difference of feature distributions in the sets. However, in Sec. 10.2.2, the warping of the neutral train set was surprising considering that the same set was used for the model training. The warp rate even increased in the second VTLN retraining iteration. A similar phenomenon was observed in (Lee and Rose, 1996), where an effect of *VTLN retraining* was studied up to 4 iterations.

To analyze a further evolution of the warping trend, the number of *VTLN retraining* periods was extended in this section. The HMM training from scratch started with 36 iterations without VTLN. Subsequently, *VTLN retraining* was applied and repeated consecutively after every 3 standard retraining iterations until a total of 120 iterations was reached. In the previous sections, similar trends in the α distributions were observed for female and male speakers when using corresponding gender dependent models, hence, the analysis presented in this section was carried out on the female data only. Evolution of the smoothed α distribution during the training is shown in Fig. 10.5. For all HMM sets, the α values represent the degree of warping of the original unwarped data (e.g., $\alpha = 1$ refers to the FB spread over 0–3200 Hz). It can be seen that in iterations 39–75, the maximum of the α distribution is reached for $\alpha = 0.9$. For higher iterations, the maximum shifts down to $\alpha = 0.85$.

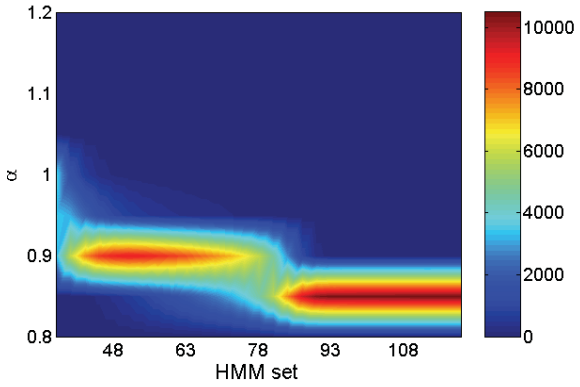

Figure 10.5: Evolution of utterance-dependent α distribution – HMM training, female train set, $F_s = 8$ kHz.

HMMs obtained from each *VTLN retraining* iteration were tested in the digit recognition task employing *utterance-dependent* and *speaker-dependent VTLN recognition*. Corresponding *utterance-*

dependent α distributions for the neutral and LE open test set are shown in Fig. 10.6. In the case of

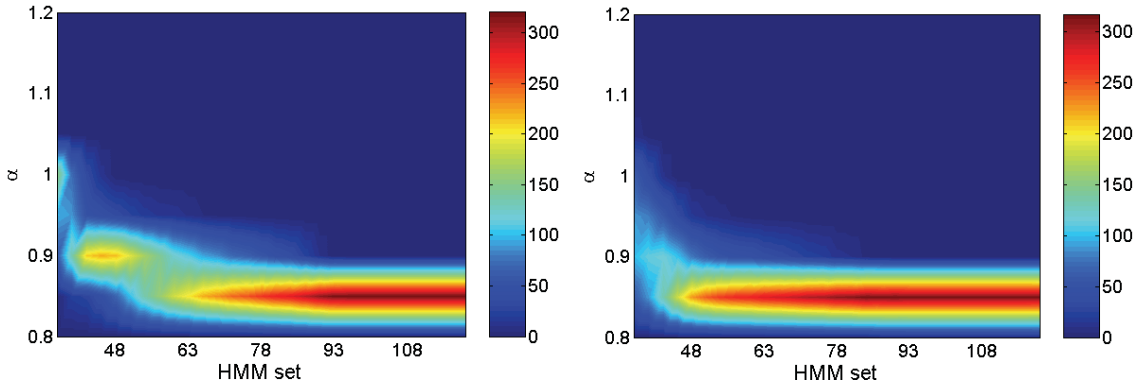

Figure 10.6: Evolution of utterance-dependent α distribution – VTLN recognition, female neutral (left figure) and LE (right figure) open test set, $F_s = 8$ kHz.

the HMMs obtained in the initial *VTLN retraining* iterations, the maxima of α distributions appear at $\alpha = 1$ for neutral set and at $\alpha = 0.9$ for LE set. With increasing number of retraining periods, the maxima of both neutral and LE α distributions shift to lower values. The trajectory is steeper for the LE set.

Up to 57 retraining iterations, the α distribution for the neutral *VTLN recognition* more or less copy the *VTLN training* distribution, which could be expected considering that both the train and test sets capture the same talking style. For higher HMM sets – up to HMM84, the test set requires more warping ($\alpha = 0.85$) than the train set ($\alpha = 0.9$). Starting from HMM84, both train and test set settle at $\alpha = 0.85$ and do not display any further shift of the distribution maxima. In the case of the LE test set, the α distribution maximum shifts soon from the initial $\alpha = 0.9$ to $\alpha = 0.85$ where it stays until HMM120. Up to HMM57, spectra of the LE data require more shrinking than the neutral test data. Starting from HMM57, the maxima of α distributions for both neutral and LE test sets occur at the similar position ($\alpha = 0.85$), although, intuitively, a degree of the LE data warping should be higher.

Since the data employed in the presented experiment were down-sampled from 16 kHz to 8 kHz and passed through the G.712 telephone filter, the frequency components above 3700 Hz are significantly suppressed, see Fig. 9.1. The warping $\alpha = 0.85$ corresponds to the FB widening from 0–3200 Hz to 0–3765 Hz. Further FB expansion introduces highly attenuated data to the highest FB bands, causing mismatch with models trained on the non-attenuated part of spectra. Frequency components below 300 Hz are also suppressed due to applying the telephone filter, however, this attenuation is present both in the training and test data and, hence, does not contribute to the models/data mismatch. Some of the low frequency components are still relevant for speech recognition, see example in Fig. 9.2, where omitting the lowest band (0–200 Hz) resulted in an increase of WER on neutral speech.

To analyze the warping trend when no telephone filter is applied, the experiment was repeated using full-band data sampled at 16 kHz. Since the broader frequency band allows for more extensive warping, the interval of α candidates for warping the initial FB of 0–3200 Hz was extended to 0.4–1.2, where $\alpha = 0.4$ corresponds to the full-band of 0–8 kHz. Number of retraining periods was increased to 174. Besides this, the training and recognition procedure were preserved identical with the previous setup. The 16 kHz α distributions for *VTLN training* and *VTLN recognition* are shown in Fig. 10.7, 10.8. Similar to the previous experiment, the data are consecutively shrunk with increasing number of retraining periods, however, the maxima of the α distributions shift down faster and do not stop at $\alpha = 0.85$. It can be seen that the maxima for the neutral test data occur at lower α's than for

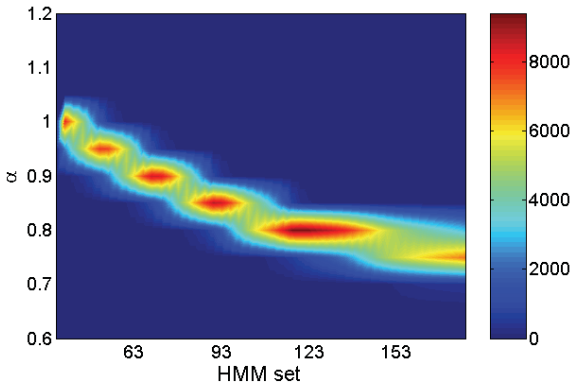

Figure 10.7: Evolution of utterance-dependent α distribution – HMM training, female train set, $F_s = 16$ kHz.

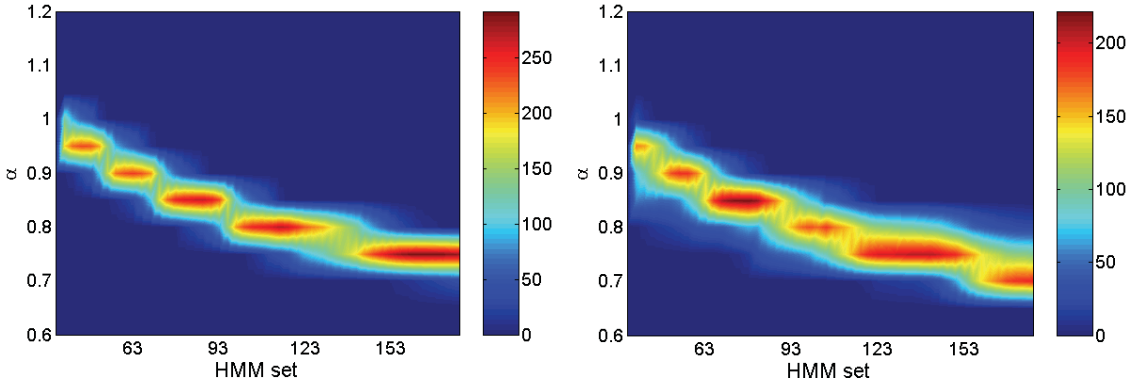

Figure 10.8: Evolution of utterance-dependent α distribution – VTLN recognition, female neutral (left figure) and LE (right figure) open test set, $F_s = 16$ kHz.

the neutral train data and that the maxima for the LE test data occur at lower α's than those of the neutral test data.

Also in (Lee and Rose, 1996), a requirement to warp data used for the model training was observed. The authors assume that if these formerly discarded portions of spectra carry useful information for recognition, the ML estimate of α is likely to be biased towards frequency shrinking which would employ originally discarded frequency components into the process of feature extraction. The experiments carried out in this section confirm this assumption. In the case of telephone speech, the FB expansion stopped when the highest FB band approached attenuated frequency components disturbing recognition. When switching to the full-band 16 kHz data, FB continuously expanded throughout all retraining iterations. The rate of expansion gradually decreased with the number of iterations. It seems that the trend of VTLN warping is driven by two factors:

- Capturing useful information lying in higher frequencies,

- Decreasing frequency resolution due to FB expansion,

The first factor assumes that a particular information useful for speech recognition has already been seen in the highest FB band for some portion of data and, hence, is partly captured in the models[3]. When performing *VTLN retraining*, the data which already contributed by this 'edge' information to the model training are likely to stay unwarped. Shrinking the remaining data will help to move the particular information to the FB area also in their case. For this reason, the mean of the α distribution during training is always at $\alpha \leqslant 1$. The second factor represents an information loss due to the decrease of frequency resolution. This factor was slowing down the FB expansion with the increasing number of retraining iterations in the 16 kHz experiment.

Performance of utterance and speaker-dependent *VTLN recognition* for models obtained from *VTLN retraining* is shown in Fig. 10.9, 10.10. It can be seen that utterance-dependent and speaker-dependent warping provide comparable results. For Lombard speech, the models from the first or second *VTLN retraining* iteration displayed the best performance. For neutral speech, HMM48 and HMM51 performed best at 8 kHz and HMM48 and HMM51 together with HMM66 and HMM69 displayed the best results at 16 kHz data. With further increasing the number of *VTLN retraining* iterations, the recognition performance tends to decrease, which can be attributed to the model over-fitting to the training data. Similar observation was reported by (Lee and Rose, 1996), where the effect of *VTLN retraining* was evaluated up to 4 iterations.

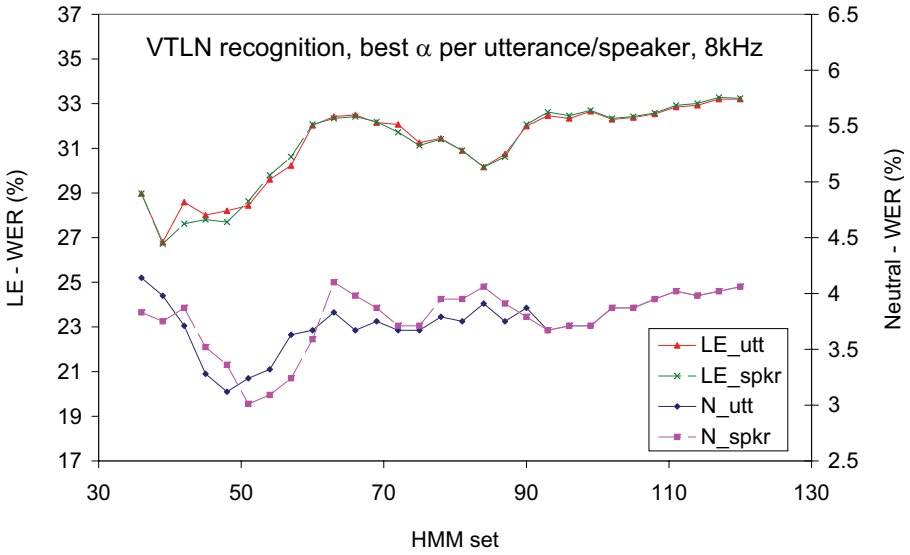

Figure 10.9: Performance of joint VTLN training and recognition, female open set, $F_s = 8$ kHz.

[3]E.g., let FB high cut-off be placed somewhere at the location of the F_4 typical occurrence. For speakers with longer VTL, F_4 will occur at lower frequencies captured by FB while for speakers with shorter VTL, F_4 will be out of FB range and thus discarded.

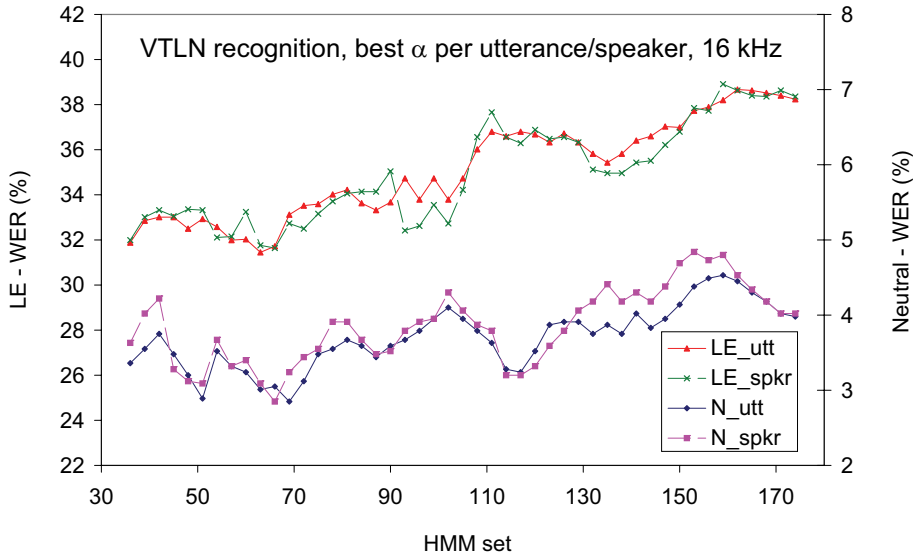

Figure 10.10: Performance of joint VTLN training and recognition, female open set, $F_s = 16$ kHz.

10.3 Formant-Driven Frequency Warping: Background

In (Eide and Gish, 1996), a parametric approach to VTLN has been proposed. Optimal warping factor for a given speaker was estimated as the ratio of the median of the speaker's F_3 estimates and the median of F_3 estimates taken from all training speakers. The formant values were extracted from the voiced frames of speech. The parametric VTLN approach was successful in improving ASR performance on neutral speech while being less computationally complex compared to ML–VTLN. In this section, a modified scheme of formant-driven warping is proposed.

10.4 Formant-Driven Frequency Warping: A Novel Approach

10.4.1 Warping Functions

Similar to ML–VTLN, (Eide and Gish, 1996) conducted warping of the original frequency axis F by a constant: $F_W = F/\alpha$. Such a frequency mapping is represented by a straight line passing through the coordinate origin of the F–F_W plane. Since formant frequencies are inversely proportional to VTL, the mapping seems to be suitable for covering the VTL variations. However, when studying differences between neutral and Lombard speech, it can be seen that formant shifts introduced by LE do not simply follow the warping by a constant. In CLSD'05, $F_{1,2}$ of the LE vowels shift considerably to higher frequencies while no consistent change in $F_{3,4}$ locations can be observed[4], see Sec. 6.5. Here, the warping by a constant could either accurately address the shifts of low formants, at a cost of introducing unwanted shifts to high formants, or it could preserve locations of high formants, disregarding the shifts of low formants. To eliminate the tradeoff between the accuracy of low and high formant mapping, new warping function is proposed in this section:

$$F_W = \alpha + \beta F. \tag{10.5}$$

[4]This suggests that the articulator configuration changes due to LE do not only introduce changes in speaker's actual VTL, but also affect individually each of the formants.

Such a frequency mapping, still being represented by a straight line, can capture both considerable frequency shifts at low frequencies and reduced shifts at high frequencies. Parameters of the transformation are estimated by regressing mean F_{1-4} values in the LE frequency–neutral frequency plain.

A 16 kHz version of the data sets from Sec. 10.2.1 was used in the presented experiment. First, distributions of F_{1-4} were estimated from all speech segments, see Fig. 10.11, and Table 10.3, 10.4.

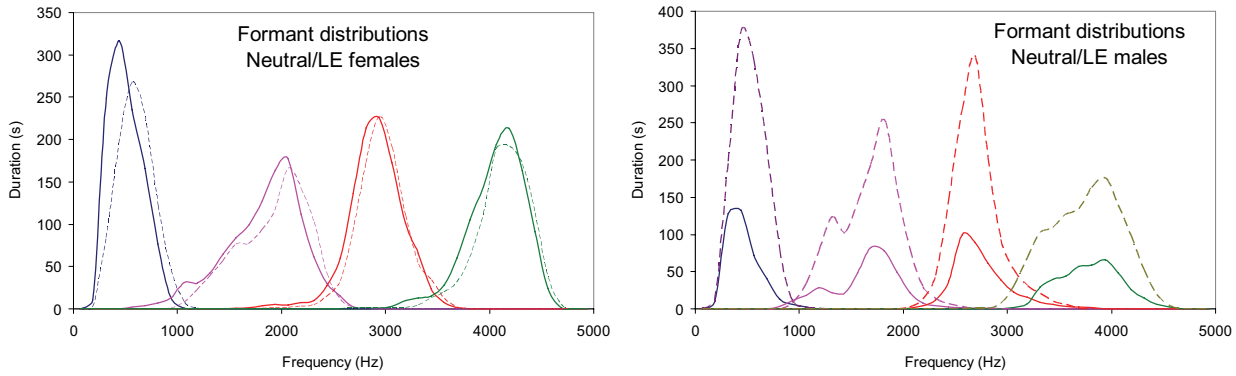

Figure 10.11: Formant distributions, neutral and LE speech, female and male utterances. Neutral speech – continuous line, Lombard speech – dashed line.

Formant		F_1	F_2	F_3	F_4
Female digits N	$\bar{F}$ (Hz)	507.0	1836.1	2904.1	4064.7
	σ (Hz)	162.7	373.3	272.9	278.8
Female digits LE	$\bar{F}$ (Hz)	598.7	1925.6	2964.9	4129.0
	σ (Hz)	169.0	336.1	241.5	261.0

Table 10.3: Female digits – means and deviations of formant distributions.

In female digits, all LE formants moved to higher frequencies, $F_{1,2}$ displaying a bigger shift than $F_{3,4}$.

Formant		F_1	F_2	F_3	F_4
Male digits N	$\bar{F}$ (Hz)	463.7	1649.8	2742.1	3785.1
	σ (Hz)	160.6	330.3	272.3	298.5
Male digits LE	$\bar{F}$ (Hz)	524.7	1670.8	2701.0	3749.5
	σ (Hz)	153.9	296.7	234.8	343.9

Table 10.4: Male digits – means and deviations of formant distributions.

In male digits, less significant changes can be observed – $F_{1,2}$ shifted up the frequency axis while $F_{3,4}$ moved slightly down. In general, typical locations of formant occurrences vary for different phonemes (see Fig. 6.3). These differences might cause an occurrence of multiple peaks in the distribution of a particular formant across phonemes. In spite of this, formant distributions collected from digit phonemes in CLSD'05 are close to their Gaussians, see Fig. 10.12, 10.13. The highest deviation from its Gaussian can be observed for F_2.

In the next step, a transfer function mapping LE formants to neutral formants was determined. Linear regression, see Eq. (4.22)–(4.24) in Sec. 4.4, was applied to the 2–D points $P_i\left(x_i, y_i\right) = P_i\left(\bar{F}_{LE,i}, \bar{F}_{N,i}\right)$, where i is the formant index, $\bar{F}_{LE,i}$ denotes the estimated mean of the i-th LE

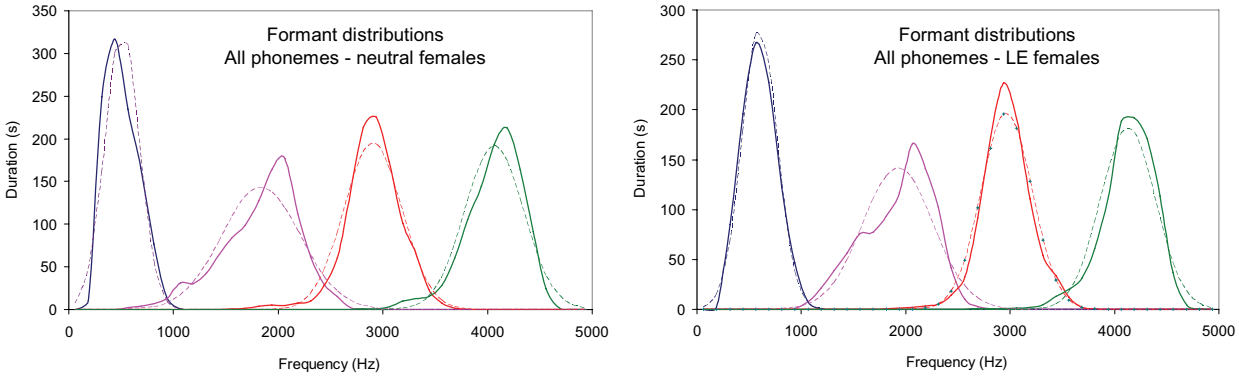

Figure 10.12: Formant distributions and their Gaussians, neutral speech, females.

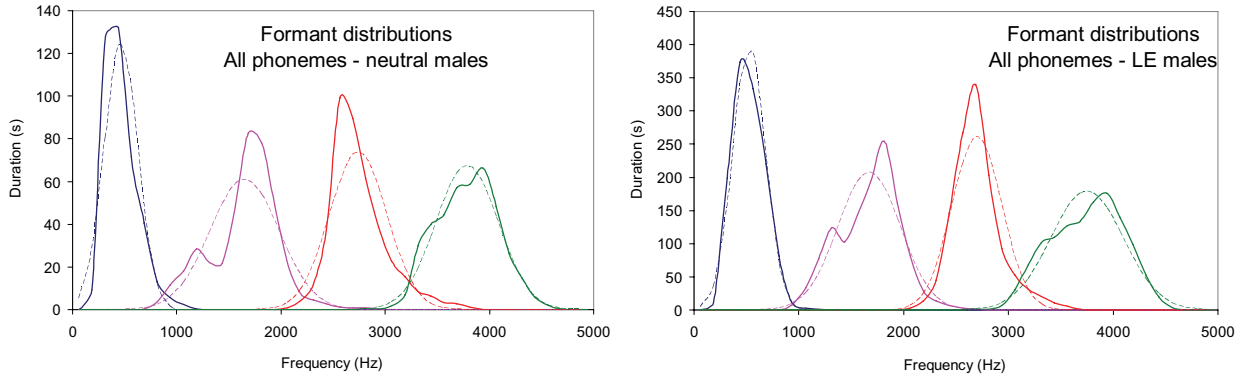

Figure 10.13: Formant distributions and their Gaussians, neutral speech, males.

formant distribution, and $\bar{F}_{N,i}$ stands for the estimated mean of the i-th neutral formant distribution. To evaluate the degree of linear relationship between x and y variables, the square of the correlation coefficient[5] was evaluated, Stephens (1998):

$$R^2 = \left(\frac{\sum_{i=1}^{N} \left(x_i - \bar{X}\right)\left(y_i - \bar{Y}\right)}{(N-1)\,\hat{\sigma}_x \hat{\sigma}_y} \right)^2 , \tag{10.6}$$

where $\bar{X}$, $\bar{Y}$ denote the LE and neutral formant sample mean, and $\hat{\sigma}_x$, $\hat{\sigma}_y$ are standard deviation estimates, see Eq. (4.13) in Sec. 4.2.1. The resulting warping functions for female and male speech are shown in Fig. 10.14, 10.15. For illustration, mean formant positions for each of the digit phonemes were also estimated by means of formant tracking and forced alignment. In the figures, each phoneme is represented by four dots standing for the four formants. The formant plots are distinguished by color and shape.

[5]The correlation coefficient R indicates a degree of linear dependence between two variables. The closer is R to 1 or -1, the stronger is the correlation between the variables. When R equals 1 or -1, the variables are linearly dependent and display increasing or decreasing relationship, respectively, Stephens (1998). The squared correlation coefficient R^2 reaches values between 0–1. The closer to 1, the more linear is the dependency between the values.

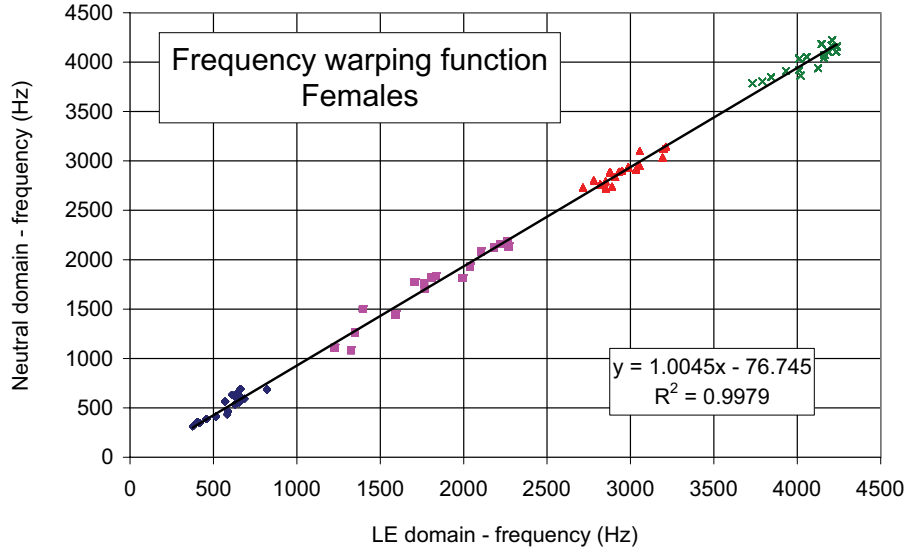

Figure 10.14: Frequency warping function, females.

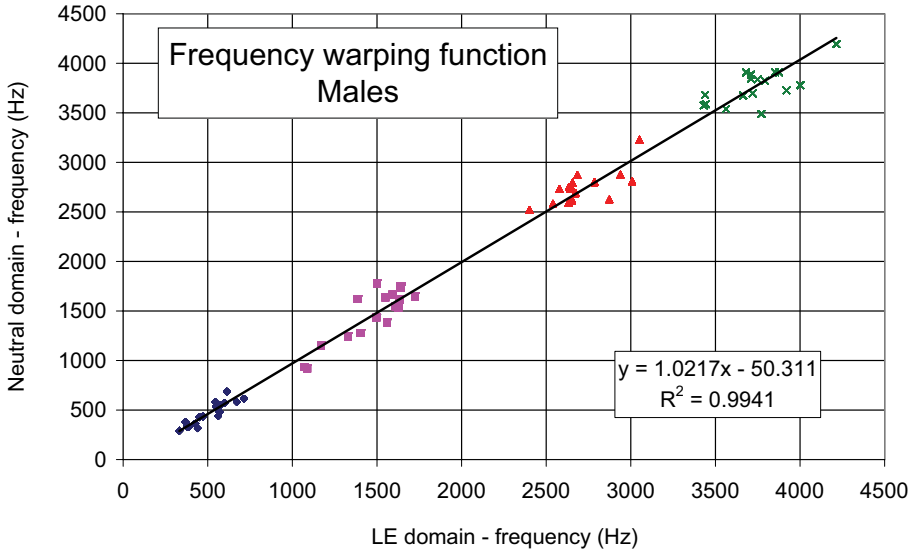

Figure 10.15: Frequency warping function, males.

10.4.2 Recognition Employing Formant-Driven Warping

Efficiency of the formant-driven frequency warping was evaluated in the ASR task. The warping functions obtained in the previous section were used for normalizing test data towards neutral, unnormalized gender-dependent acoustic models. To allow for performance comparison with the *VTLN recognition* experiment, (see Sec. 10.2.1, where data normalization was applied also only in the recognition stage), the data were down-sampled to 8 kHz. Similarly as in *VTLN recogniton*, feature extraction for the unnormalized model training employed PLP features with 20-bands rectangular FB spread over 0–3200 Hz (baseline FB). The recognizer was trained in 36 iterations using progressive

mixture splitting. Filter banks for the test data extraction were derived from the baseline FB by applying gender-dependent warping functions to its endpoints, see Table 10.5.

Filter bank	F_{start} (Hz)	F_{stop} (Hz)	# Bands
Baseline	0	3200	20
Warped F	76	3262	20
Warped M	49	3181	20

Table 10.5: Baseline and gender-dependent, warped filter banks.

Recognition performance reached for the unwarped data, 'baseline bank', and warped data, 'warped bank', is shown in Table 10.6. It can be seen that for both female and male Lombard speech, frequency warping significantly reduces WER over the baseline. Surprisingly, when applying warping also to the neutral utterances, the performance is not significantly corrupted – statistically insignificant WER changes are displayed for both females and males. Compared to the *VTLN recognition* experiment, see Table 10.1 in Sec. 10.2.1, a more significant performance improvement can be observed in the case of the formant-driven warping[6]. However, it must be emphasized that while the warping in *VTLN recognition* was performed on the utterance level, in the case of the formant-drive approach, the warping functions were estimated using the whole open set. For the purpose of the real-world application, the warping function could be determined from the train data and later updated for estimates from the incoming test utterances.

Set		Females		Males	
		Neutral	LE	Neutral	LE
# Digits		2560	2560	1423	6303
WER (%)	Baseline bank	4.2 (3.4–5.0)	35.1 (33.3–37.0)	2.2 (1.4–2.9)	23.2 (22.1–24.2)
	Warped bank	4.4 (3.6–5.2)	23.4 (21.8–25.0)	1.8 (1.1–2.4)	15.7 (14.8–16.6)

Table 10.6: Recognition employing formant-driven warping, HMM36. Mean values followed by 95% confidence intervals in parentheses.

10.5 Conclusions

Two methods normalizing LE formant structure towards neutral were proposed in this chapter – modified maximum likelihood VTLN (ML–VTLN) and formant-driven frequency warping.

In the standard ML–VTLN recognition, utterance transcriptions are obtained by decoding the unwarped data and, subsequently, used for estimation of warping factors. To address the high transcription error rate introduced by LE, a new method of ML–VTLN considering multiple transcription hypotheses was proposed. Next, effectiveness of utterance and speaker-dependent warping in VTLN training was compared. Both methods displayed similar performance, suggesting a use of utterance-dependent warping due to its lower complexity. A considerable improvement of the Lombard speech recognition was reached both by a system employing ML–VTLN training/recognition and by a system employing just ML–VTLN recognition. Both systems displayed comparable performance in most of

[6]Note that baseline WERs for the two experiments slightly differ due to different HMM sets used, *VTLN recognition* employed HMM46 while *formant-driven warping recognition* employed HMM36.

the scenarios considered. In the case of male Lombard speech, the first system performed slightly better. Finally, development of the warping trend with increasing number of VTLN retraining iterations was studied. Accuracy of the VTLN-normalized acoustic models improved up to approximately 10 VTLN retraining iterations, with further iterations the models started to overfit the training data.

Formant-driven warping proposed in this chapter employs a linear mapping function derived from the distribution means of the first four Lombard and neutral formants. The new mapping function allows for more accurate modeling of the low and high formant shifts compared to the warping by a constant used in the literature. In the pilot study presented here, parameters of the warping function were estimated from the formant distributions across the whole test set. Applying the formant-driven warping in the recognition stage provided substantial improvement of the recognition accuracy over the baseline, outperforming also the ML-VTLN system.

Similarly as in the case of ML–VTLN, real-world applications would require the formant-driven algorithm to be able to react at the level of the actual incoming utterances. Here, the warping function could be updated for the estimates from the actual utterance. Evaluation of such an approach is a matter of further study. Formant-driven warping is computationally less demanding than ML–VTLN, which requires multiple utterance decoding for the set of warping candidates[7]. On the other hand, the formant-driven approach relies on a formant tracking, which is easily deteriorated by the presence of background noise.

None of the frequency warping approaches considered in this chapter compensates for the narrowing of formant bandwidths due to LE, see Sec. 6.5. In fact, shrinking the LE formant structure towards neutral further decreases the LE formant bandwidths. It may be assumed that a transformation addressing both changes in formant locations and bandwidths will increase effectiveness of the LE normalization.

[7]Recently, possibilities of reducing the complexity of ML–VTLN have been studied. In (Faria and Gelbart, 2005), based on the observed correlation between F_0 and formant shifts, the optimal warping factor was estimated directly from F_0.

Chapter 11

Two-Stage Recognition System (TSR)

A substantial improvement in recognition performance can be reached by accommodating an ASR system to the actual conditions affecting the processed speech signal. Unfortunately, as a consequence, such condition-specific adjustments usually result in a reduction of the system's resistance to condition changes. Several studies have attempted to address this performance/robustness trade-off by building a recognizer of a set of condition-specific (CS) subsystems. In (Xu *et al.*, 1996), multiple HMM channels were employed to model separately speech and noise signal components. A similar HMM structure was used in (Womack and Hansen, 1999), see Sec. 3.5, where each of the HMM channels comprised stress-specific acoustic models. During utterance decoding, the multi-channel HMM allowed for stress assessment and switching between the stress-specific models on the HMM-state level.

Another concept has been proposed in (Womack and Hansen, 1996b). Here, for each incoming utterance the actual condition was first assessed by a condition classifier, and subsequently, the utterance was passed to the corresponding CS recognizer[1]. This approach does not allow for switching between the CS models on the sub-word level, on the other hand, computational demands are considerably reduced here since the complexity of the utterance decoding in the CS single-channel HMM recognizer is lower compared to the multi-channel HMMs.

Both the multi-channel HMM and 'condition classifier/CS recognizers' approaches mentioned above assume that the conditions likely to occur are known already during the system design and a sufficient amount of training data is available per each of these conditions. In (Bou-Ghazale and Hansen, 1998), a lack of training samples per scenario was compensated for by generating synthetic CS speech from neutral samples using perturbation models, see Sec. 3.5.

In Chap. 9, features obtained from the data-driven design provided a significant improvement of the Lombard speech recognition when using neutral-trained acoustic models. Increasing the feature performance on Lombard speech was accompanied by a performance drop on neutral speech. In this chapter, two front-ends observed to provide superior performance on neutral and Lombard speech, respectively, are used in the design of a two-stage system (TSR) for neutral/LE speech recognition[2]. Following the 'condition classifier/CS recognizers' architecture, in the first stage of TSR, the utterances are classified based on talking style (neutral/LE) and in the second stage, they are passed to the corresponding CS recognizer. Unlike in the case of the previously mentioned studies, the CS recognizers require only neutral speech data for training.

This chapter is organized as follows. First, the architecture of the neutral/LE classifiers used in the TSR experiments is described. Subsequently, neutral and LE distributions of spectral slope are analyzed. Spectral band limitation is used to reduce the overlap of the neutral/LE slope distributions. Performance of a set of candidate features in the neutral/LE classification is evaluated. Several

[1] An alternative scheme can be found in (Womack and Hansen, 1995a), where the output of the set of stress-dependent recognizers was weighted by a condition classifier.

[2] The design of TSR was published in (Bořil *et al.*, 2007).

candidate subsets are compared and the most efficient one yields the final classification feature vector (CFV) used for training ANN and Gaussian ML TSR classifiers. Finally, the TSR is constructed, employing the neutral/LE classifier and two neutral/LE-specific recognizers.

11.1 Classification of Neutral/LE Speech

In the following sections, the CFV for gender/phonetic content-independent neutral/LE speech classification is searched. An overview of the literature on this topic has been provided in Sec. 3.4. Significant differences in vocal intensity, F_0 means and standard deviations, $F_{1,2}$ means and bandwidths, and vowel durations were observed for neutral and Lombard speech in CLSD'05 (see Chap. 6). Since formant and duration characteristics are heavily dependent on the phonetic content of the actual utterance, they were excluded from the set of CFV candidates[3]. On the other hand, the set was extended for the spectral slope of voiced speech which has been reported in previous studies to be a reliable cue for the talking style assessment (see Sec. 3.2.4).

Neutral and Lombard digits and phonetically rich sentences from 8 female and 7 male CLSD'05 speakers formed the CFV development data set. The development set was used for analyzing neutral/LE discriminability provided by feature distributions and for training the classifiers. The open test set comprised digits and phonetically rich sentences uttered by 4 female and 4 male CLSD'05 speakers (speakers are separate from the development set)[4] and was used for the classifier open testing. The data were all in 16 kHz sample format (i.e., 8 kHz bandwidth).

11.1.1 Gaussian ML Classifier

A Gaussian ML classifier was used in the following experiments which consisted of two n-dimensional PDFs modeling CFV component distributions in neutral and LE speech. Each PDF dimension contained a single Gaussian and the number of dimensions n corresponded to the number of CFV components. Mean vectors and covariance matrices of *neutral PDF* and *LE PDF* were estimated from the training data. Bayesian hypothesis testing was applied during classification. For each CFV extracted at the utterance level (one CFV per utterance), two hypotheses were tested: H_0 – the utterance comprises neutral speech, H_1 – the utterance comprises Lombard speech. Given the CFV observation $\boldsymbol{o_i}$ extracted from the utterance i, a likelihood ratio was calculated:

$$\lambda = \frac{P\left(\boldsymbol{o_i}\,|H_1\right)}{P\left(\boldsymbol{o_i}\,|H_0\right)}, \tag{11.1}$$

where $P\left(\boldsymbol{o_i}\,|H_0\right)$ and $P\left(\boldsymbol{o_i}\,|H_1\right)$ are conditional probabilities evaluated by applying (2.8) (Sec. 2.2) to the *neutral PDF* and *LE PDF*, respectively. For $\lambda < 0.5$[5], H_0 was accepted and the utterance was classified as neutral, otherwise, H_0 was rejected and the utterance was classified as LE.

11.1.2 ANN Classifier

As an alternative to the Gaussian ML classifier, a fully-connected three-layer feed-forward Multi-Layer Perceptron (MLP) was used in TSR to estimate the posterior probabilities of utterances being

[3]The inefficiency of using mean vowel $F_{1,2}$ locations as features in the classification task was experimentally verified by (Zhou *et al.*, 1998).

[4]The assignment of female speakers to the development or open set was preserved as in Sec. 9.1 to assure that none of the open test data used for evaluating the TSR performance in the end of this chapter were also employed in the data-driven feature design or in the CFV/classifier design.

[5]Adjusting the threshold for H_0 rejection can be used to optimize the ratio of false acceptances and rejections, (Zhou *et al.*, 1998). In the presented experiments, the threshold was kept constant.

either neutral or LE. The number of neurons in the MLP's[6] input layer was set equal to the number of components in the CFV. The hidden layer contained 3000 neurons and the output layer 2 neurons. Each neuron in the hidden layer and output layer implemented activation function F according to Fig. 11.1 where $A_1, A_2, \ldots, A_N$ are the incoming neural activations, $W_{1j}, W_{2j}, \ldots, W_{Nj}$ are the incoming

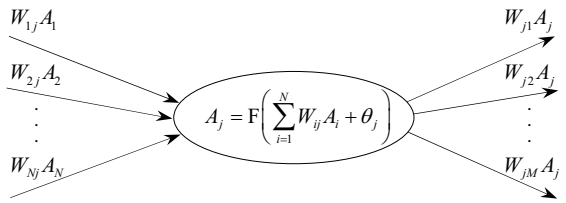

Figure 11.1: Multi-Layer Perceptron – activation function, after (Stergiou and Siganos, 1996).

connection weights, θ_j is the bias, A_j is the output activation, and $W_{j1}, W_{j2}, \ldots, W_{jM}$ are the output connection weights. In the hidden layer, a sigmoid function was used as the activation function:

$$\mathrm{sig}\,(q_j) = \frac{1}{1 + e^{-q_j}}, \tag{11.2}$$

where q_j is the biased sum of weighted inputs to the j-th neuron:

$$q_j = \sum_{i=1}^{N} W_{ij}A_i + \theta_j. \tag{11.3}$$

The sigmoid function maps the input values to the interval (0, 1). For input values out of the range (-1, 1) the output becomes insensitive, reaching values very close to 0 for negative input values and values very close to 1 for positive inputs. The function is smooth and easily differentiable, which are important assumptions for applying the traditional gradient-based MLP training procedure called *back propagation*, (Rabiner and Juang, 1993). In the output layer, a softmax activation function was used:

$$\mathrm{softmax}\,(q_j) = \frac{e^{q_j}}{\sum_{i=1}^{M} e^{q_i}}, \tag{11.4}$$

where q_i are the weighted input sums in the output layer neurons (see Eq. (11.3), substitute q_i for q_j). A softmax function assures that the MLP outputs will reach values interpretable as posterior probabilities (i.e., each lying in the interval $\langle 0, 1\rangle$ and all together summing up to 1).

The training procedure implemented in *QuickNet* originates from *back propagation* where the partial derivatives of the MLP's output error (difference between the training sample target value and the MLP's output) are used to adjust the MLP weights. The weight adjustment step, called the *learning rate*, is chosen as a compromise between the speed, convergence, and precision of MLP learning.

For the purposes of MLP training, a small set of samples (cross-validation set – the CV set) was excluded from the training set. In each training period, the entire training set was processed and MLP weights were adjusted. Subsequently, the actual MLP performance was evaluated on the CV set. In the initial retraining periods, the *learning rate* was set to a relatively high value. After the performance improvement between two consecutive periods decreased below 0.5 %, the *learning rate* was halved before each subsequent period to increase precision of the local optimum search. The

[6]The MLP framework was implemented using the ICSI *QuickNet* suite, (QuickNet, 2007).

training process was terminated when the performance improvement decreased again below 0.5 % between consecutive training periods.

The described adaptive training scheme is used to protect the MLP from over-training, (QuickNet, 2007). When further retraining increases the MLP performance on the training set but at the same time reduces the performance on the CV set, it can be assumed that the MLP starts to overfit the training data and has reduced generalization properties. To prevent from this from happening, the training procedure is terminated once the performance improvements on the CV set decrease below a chosen threshold[7].

In the experiments presented in the following sections, MLP was trained against hard targets – the required MLP outputs were either 0 or 1 (either neutral or LE). 90 % of the development data were used for MLP training and 10 % for CV.

11.1.3 Exploiting Spectral Slope

As discussed in Sec. 3.2.4, spectral slope of the glottal pulses provide an important cue to the talking style classification. Spectral slope of voiced speech is determined by combined contribution of glottal pulse spectrum and transfer function of the lip radiation. Since the lip radiation can be considered to have a constant slope (+6 dB/oct), see Sec. 3.1, the slope of the glottal pulse spectrum is directly proportional to the slope of the voiced speech spectrum, (Childers and Lee, 1991). In this section, the neutral/LE discriminability provided by the spectral slope of voiced speech is studied. Spectral slopes were obtained by fitting a regression line to the amplitude spectra of voiced segments of speech, see Sec. 4.4, following (Summers *et al.*, 1988). In the initial step, spectral slopes were extracted from the full-band spectrum, i.e., from the frequency interval of 0–8 kHz.

Vowel Slopes

First, mean spectral slopes in vowels were analyzed. Time boundaries of the development set phonemes were estimated by means of forced alignment. These boundaries were used to determine endpoints of the variable length segments used for the spectral slope extraction. Before conducting the actual extraction, each segment was weighted by a Hamming window. An example of the spectral slope of a single realization of the female vowel /a/ is shown in Fig. 11.2. Mean slopes of five Czech vowels /a/, /e/, /i/, /o/, /u/ were estimated from the development set (see Table 11.1, 11.2). Here, $\#N$, $\#LE$ denote the number of neutral and LE phoneme occurrences in the development set, respectively. It can be seen that spectral slope differs across vowels and is in general steeper for female voices in both neutral and LE utterances. The slope is less steep in LE speech, confirming observations made in previous studies (see Sec. 3.2.4). The slope flattening results from migration of the glottal pulse spectral energy to higher frequencies.

Voiced Segment Slopes

The task of the classifier employed in the TSR system is to reliably discriminate neutral and Lombard speech without information regarding phonetic content of the actual utterance or gender of the speaker. To evaluate the suitability of spectral slope for participation in CFV, its phoneme-independent distributions in neutral and Lombard data are studied in this section, followed by analysis of the gender-independent slope distributions in the subsequent section.

Neutral and LE phoneme-independent distributions were obtained by extracting spectral slope from all voiced segments of neutral and Lombard speech, respectively. Locations of voiced segments

[7]The threshold is experimentally set to limit the number of retraining iterations to the extent where further retraining is not expected to provide significant MLP performance gains. In the present experiments, the default *QuickNet* threshold of 0.5 % was used.

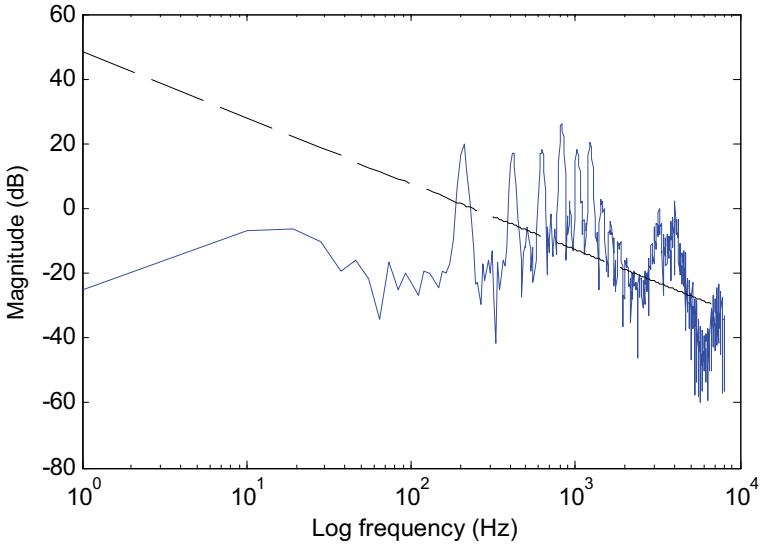

Figure 11.2: Example – single realization of female vowel /a/, amplitude spectrum and corresponding spectral slope (−6.09 dB/oct).

were estimated based on the output of the pitch tracker. Every segment yielding a positive F_0 value was considered voiced. Two variants of spectral slope extraction were used. In the first variant, the slope was extracted from the frequency band of 0–8000 Hz (full-band slope), similarly as in the previous section. In the second variant, the band for extraction was limited to 60–8000 Hz. The latter setup was motivated by the assumption that in voiced speech, the F_0 subharmonics contain noise rather than speech related information and, hence, do not contribute to the neutral/LE classification. Means and standard deviations of the slopes obtained from both extraction setups are shown in Table 11.3. It can be seen that even when extracted across all voiced segments, the mean values of spectral slope significantly differ for neutral and Lombard speech. In the case of the band-limited setup, the slopes are steeper by omitting the initial low-energy portion of the spectra. Neutral/LE discriminability provided by full-band slopes and '60–8000 Hz' slopes will be analyzed in the following subsection.

Gender-Independent Slopes

In the next step, gender/phoneme-content independent distributions of spectral slope were analyzed. The male and female neutral development sets were merged into a single neutral set, similarly, the LE sets were merged into a single LE set. Several variants of slope extraction were examined. To allow for analysis of the neutral/LE discriminability provided by the particular slope extraction setup, the neutral and LE slope distributions were constructed and normalized to an equivalent surface. The area bounded by the overlap of the two distributions represents a portion of data which cannot be unambiguously assigned to one of the classes using the given slope extraction scheme.

Neutral/LE distribution overlaps obtained from slope extraction employing two types of weighting windows and various frequency band-limiting were evaluated:

- Band 0–8000 Hz (rectangular/Hamming window) – full-band spectrum,
- Band 60–8000 Hz – non-speech portion of spectrum excluded,
- Band 60–5000 Hz – area of F_0–F_4 occurrence,
- Band 1000–5000 Hz – area of F_2–F_4 occurrence,

Vowel	Neutral				LE			
	# N	T (s)	Slope (dB/oct)	σ (dB/oct)	# LE	T (s)	Slope (dB/oct)	σ (dB/oct)
/a/	454	69.03	-6.8 (-6.9; -6.7)	1.13	350	73.40	-3.2 (-3.4; -3.0)	1.78
/e/	1064	69.33	-5.6 (-5.7; -5.6)	1.06	840	83.55	-3.1 (-3.2; -3.0)	1.41
/i/	509	58.92	-5.0 (-5.1; -4.9)	1.15	405	64.26	-2.5 (-2.7; -2.3)	1.75
/o/	120	9.14	-8.0 (-8.1; -7.8)	0.91	90	13.44	-4.5 (-4.8; -4.2)	1.61
/u/	102	5.73	-6.1 (-6.3; -6.0)	0.77	53	2.09	-3.9 (-4.4; -3.5)	1.55

Table 11.1: Mean spectral slopes of female digit vowels, band 0–8 kHz. Mean values followed by 95% confidence intervals in parentheses.

Vowel	Neutral				LE			
	# N	T (s)	Slope (dB/oct)	σ (dB/oct)	# LE	T (s)	Slope (dB/oct)	σ (dB/oct)
/a/	162	25.99	-8.0 (-8.2; -7.8)	1.11	484	101.14	-4.9 (-5.0; -4.7)	1.78
/e/	414	25.37	-7.1 (-7.2; -7.0)	1.14	1072	102.63	-3.9 (-4.0; -3.8)	1.53
/i/	260	28.71	-6.5 (-6.7; -6.4)	1.15	528	88.67	-3.8 (-3.9; -3.6)	2.00
/o/	54	3.29	-8.2 (-8.5; -7.9)	1.06	129	14.35	-5.7 (-6.0; -5.5)	1.45
/u/	43	2.14	-7.1 (-7.4; -6.9)	0.96	97	3.78	-5.0 (-5.1; -4.8)	0.92

Table 11.2: Mean spectral slopes of male digit vowels, band 0–8 kHz. Mean values followed by 95% confidence intervals in parentheses.

- Band 0–1000 Hz – formants starting from F_2 excluded,
- Band 60–1000 Hz – only F_0–F_1.

Frequency bands in the first two setups were preserved as in the previous section. The band 60–5000 Hz in the third setup was chosen to exclude either F_0 subharmonics and highly attenuated frequency components corresponding to F_5 and higher (see Fig. 10.12, 10.13 in Sec. 10.4.1). Bands in the remaining setups were chosen to allow for separate analysis of F_2–F_4 and F_0–F_1 component contributions in neutral/LE discrimination.

In the case of the first extraction setup, normalized full-band slope distributions displayed an overlap of 28.06 % for a Hamming window and 32.24 % for a rectangular window. The higher distribution overlap observed in the latter case can be attributed to the increased spectral blending typical for signal weighting by a rectangular window. Based on this result, a Hamming window was used for all remaining experiments. Overlaps of the normalized slope distribution are listed in Table 11.4 in the row 'M+F'. For comparison, overlaps of the corresponding gender-dependent slope distributions are also provided. The normalized gender-independent slope distributions are depicted in Fig. 11.3.

Set		Neutral				LE			
		# N	T (s)	Slope (dB/oct)	σ (dB/oct)	# LE	T (s)	Slope (dB/oct)	σ (dB/oct)
0–8000 Hz	M	2587	618	-7.42 (-7.48; -7.36)	1.53	3532	1114	-5.32 (-5.37; -5.27)	1.55
	F	5558	1544	-6.15 (-6.18; -6.12)	1.30	5030	1926	-3.91 (-3.96; -3.86)	1.77
60–8000 Hz	M	2587	618	-8.68 (-8.75; -8.61)	1.82	3532	1114	-6.44 (-6.50; -6.38)	1.82
	F	5558	1544	-7.18 (-7.22; -7.14)	1.57	5030	1926	-4.89 (-4.95; -4.83)	2.03

Table 11.3: Comparing full-band and 60–8000 Hz slopes. Mean values followed by 95% confidence intervals in parentheses.

Set	Neutral – LE distribution overlap (%)					
	0–8000 Hz	60–8000 Hz	60–5000 Hz	1k–5k Hz	0–1000 Hz	60–1000 Hz
M	26.00	28.13	29.47	100.00	27.81	27.96
F	26.20	28.95	16.76	100.00	25.75	22.18
M+F	28.06	30.48	29.49	100.00	27.54	26.00

Table 11.4: Slope distribution overlaps for various extraction schemes, Hamming window.

Limiting the full-band to 60–8000 Hz resulted in an increase of the distribution overlap. Since this limiting was supposed to exclude non-speech components from the slope extraction scheme, this result is somewhat surprising. A possible explanation can be as follows. During the voiced speech production, besides vocal fold vibration, a noise component can also participate in the vocal tract excitation (see Sec. 3.1), causing an occurrence of speech production-related components at F_0 subharmonics. In such a case, the subharmonics may carry information contributing to the neutral/LE discrimination. Another reason for this result can lie in the inaccuracy of voiced segment boundary estimation. The estimated voiced regions could also contain portions of unvoiced speech which have significantly different slopes, thereby affecting the spectral slope distributions.

For the band 1000–5000 Hz, the neutral and LE slope distributions overlap completely, indicating that spectral slope does not provide any neutral/LE discriminability in this case. The last two setups represent the cases when this 'inefficient' frequency band is excluded from the extraction scheme. As could have been expected, both of them provide the lowest distribution overlaps compared to the other setups. The best discriminability is displayed by the spectral slope extracted from the band of 60–1000 Hz, yielding an overlap of 26 %.

11.1.4 Classification Feature Vector

Based on the analyses conducted in Chap. 6 and Sec. 11.1.3, the following features were chosen to participate in the set of CFV candidates:

- *SNR* – mean SNR (dB) per utterance, proportional to vocal intensity (see Sec. 6.3),
- F_0 – mean F_0 per utterance. Two frequency representations: (1) linear frequency scale, F_{0Hz} (Hz), (2) logarithmic frequency scale – cents of a semitone, $F_{0\%}$ (%), defined:

$$F_{0\%} = 1200 \log_2 \frac{F_0}{F_{ref}} \ (\%)\,, \tag{11.5}$$

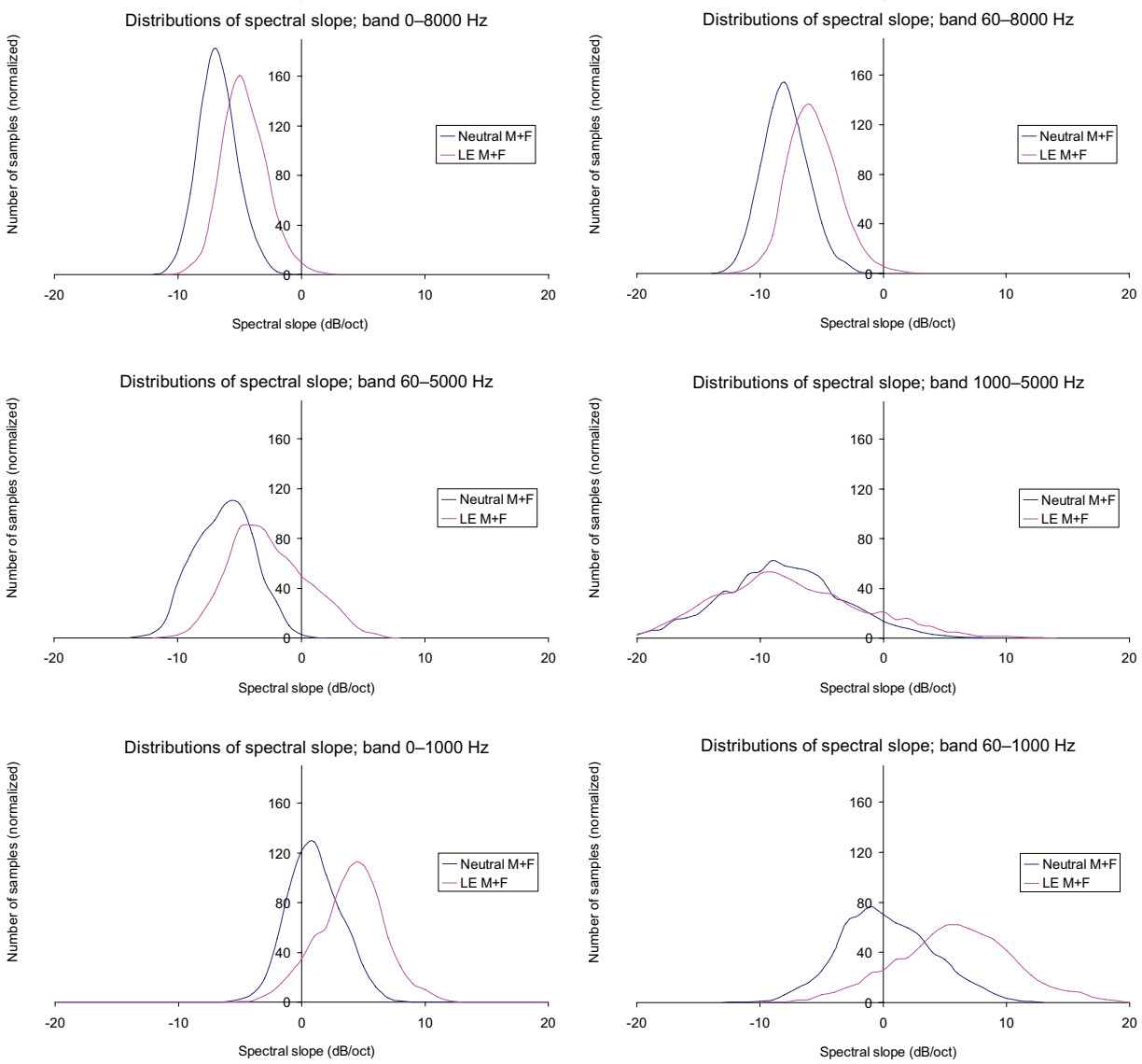

Figure 11.3: Normalized slope distributions extracted for various frequency bands.

where F_{ref} is the reference frequency (chosen $F_{ref} = 60$ Hz) and $F_{0\%}$ is the frequency difference between F_0 and F_{ref} in cents of a semitone,

- σF_0 – standard deviation of F_0 per utterance. Two representations: (1) linear frequency scale – σF_{0Hz} (Hz), (2) logarithmic frequency scale – the *tonal deviation* $\sigma F_{0\%}$ (%), deviation of F_0 represented in semitone cents, defined similarly to $F_{0\%}$,

- *Spectral slope* – mean spectral slope per utterance, extracted from the spectral band of 60–1000 Hz, $Slope_{60-1kHz}$ (see Sec. 11.1.3).

Mean values of F_0, *SNR*, and *spectral slope*, as well as σF_0 were extracted within utterances, yielding a single feature vector per each utterance. Besides the linear representation of F_0 and σF_0, their

semitone variants $F_{0\%}$, $\sigma F_{0\%}$ were also included in the candidate set, representing pitch measures inspired by human pitch perception[8].

To evaluate neutral/LE discriminability provided by the particular CFV candidates, each was consecutively employed in the training and testing of a single-feature MLP classifier. For this purpose, the development set described in the beginning of Sec. 11.1 was split into a training set (2202 utterances) and CV set (270 utterances). Both sets comprised utterances from the same male and female speakers. Performances of the single-feature classifiers are shown in Table 11.5. UER stands for the Utterance Error Rate – the ratio of the number of incorrectly classified neutral/LE utterances to the number of all classified utterances. The *Train* row represents the closed test performances and the CV row represents the open test performances. The best neutral/LE discriminability was displayed by SNR_{dB},

Set		SNR_{dB}	F_{0Hz}	σF_{0Hz}	$F_{0\%}$	$\sigma F_{0\%}$	$Slope_{0\text{-}8kHz}$	$Slope_{60\text{-}1kHz}$
UER (%)	Train (# 2202)	10.9 (9.6–12.2)	23.8 (22.0–25.5)	25.8 (23.9–27.6)	25.5 (23.7–27.3)	36.3 (34.3–38.3)	24.0 (22.2–25.8)	19.3 (17.6–20.9)
	CV (# 270)	12.2 (8.3–16.1)	18.5 (13.9–23.2)	25.6 (20.4–30.8)	24.8 (19.7–30.0)	31.1 (25.6–36.6)	20.4 (15.6–25.2)	18.2 (13.6–22.7)

Table 11.5: Efficiency of single-feature trained MLP classifier, merged male and female development sets. Mean values followed by 95% confidence intervals in parentheses. Gender-independent task.

followed by $Slope_{60-1kHz}$. For a comparison, also $Slope_{0-8kHz}$ was tested. Confirming the observations from the analysis of distribution overlaps, $Slope_{60-1kHz}$ outperforms the full-band spectral slope. The semitone representations of F_0 and σF_0 provided lower neutral/LE discriminability than their linear frequency equivalents. The performance drop in the case of $F_{0\%}$ seems to result from the logarithmic compression of the frequency axis which reduces difference between mean neutral and LE pitch. The *tonal variation* $\sigma F_{0\%}$ was outperformed by σF_{0Hz}, suggesting that tonal deviations within the same talking style differ for females and males. Based on these results, the following features were chosen to form the CFV set: SNR_{dB}, F_{0Hz}, σF_{0Hz}, $Slope_{60-1kHz}$. Normalized distributions of the CFV features are shown in Fig. 11.4.

Before normalizing the distributions, occurrences of the features extracted on the utterance level were weighted by the actual utterance lengths. In the plots on the right side, single-feature MLP posterior probabilities are shown, where $Pr\,(N)$ and $Pr\,(LE)$ are probabilities that the feature comes from neutral or Lombard utterance, respectively. It can be noted that overlaps of the normalized feature distributions shown in Fig. 11.4 do not simply correspond to the classification performances reached by the single-feature MLP classifier in Table 11.5. This is caused by the fact that the feature occurrences seen by the MLP during training and testing are proportional to the number of utterances generating the feature values (the MLP was trained to perform classification at the utterance level, that is, each training utterance was represented by a single feature vector during MLP training), while the depicted distributions are proportional to the number of occurrences weighted by the total duration of the utterances[9].

[8]Due to the logarithmic characteristic of human pitch perception, the F_0 variations in male and female speech that would be perceived as equivalent, i.e., variations by the same number of semitones, differ when represented in the linear scale (Hz). In particular, a semitone step in female speech typically corresponds to a bigger change in F_{0Hz} than in males as the female mean F_0 occurs at higher frequencies, see Sec. 6.4. If males and females produced similar or close *tonal deviation* in neutral speech and LE speech, respectively, $\sigma F_{0\%}$ would represent a gender-independent variant of σF_{0Hz}.

[9]Moreover, MLP models prior probability of each class (the probability of occurrence of a sample from the given class) while the normalized distributions discard this information. However, the development set used in the presented experiments was well balanced, comprising *1238 neutral* utterances and *1234 LE* utterances, hence, the priors can be considered equal.

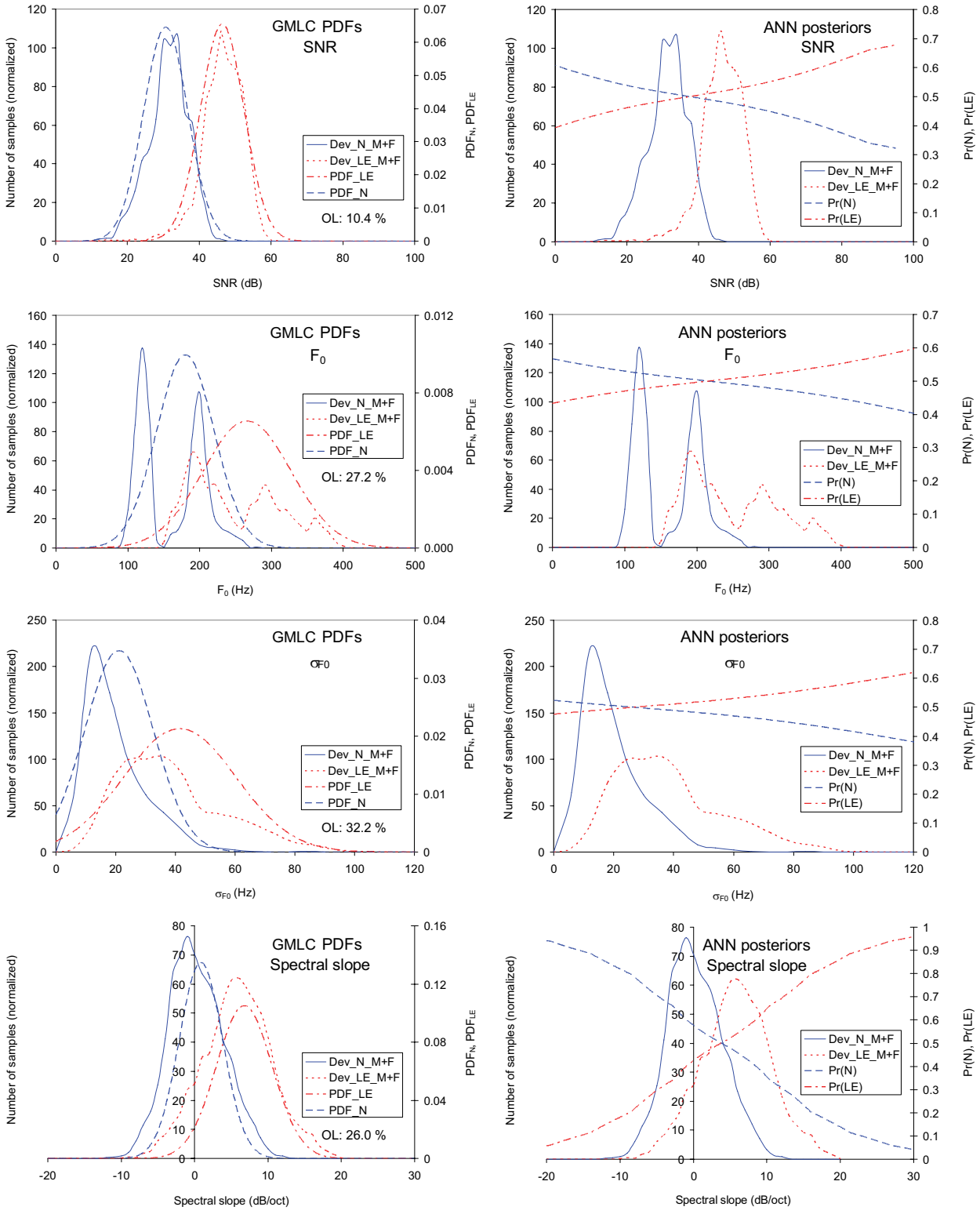

Figure 11.4: Normalized distributions of CFV features, merged male and female development sets. OL – distribution overlaps. Dashed and dash-dotted plots: left – DM–GMLC PDFs, right – single-feature ANN (MLP) posteriors (gender-independent classification).

Subsequently, MLP classifier and Gaussian ML classifier (GMLC) were trained on the development set parameterized by CFV. The development set comprised both male and female utterances. Two variants of the GMLC were trained. The first variant comprised PDFs employing full covariance matrices (FM–GMLC), the second variant employed diagonal covariance matrices (DM–GMLC). Means and variances of DM–GMLC PDFs are shown in Fig. 11.4. PDFs employing DM are frequently used in ASR systems for their computational efficiency. When using DMs, it is assumed that the modeled feature vector components are independent and, hence, their variances (forming the diagonal of the matrix) define completely the covariance matrix while the non-diagonal components, covariances, are equal to zero. If the CFV components are not completely independent, the DM PDFs will provide less accurate models of the data sets compared to FM PDFs.

Performance of MLP, DM–GMLC, and FM–GMLC was tested in the closed and open tests. The closed tests were carried out on the development set which was used for training the classifiers. The open tests employed open sets described in the beginning of Sec. 11.1. In the case of MLP, the development set was further split into the training set, representing the data used for training the MLP parameters, and the CV set, which was used to control convergence of the MLP's training. Means and standard deviations of utterance durations in the development set and open test set are shown in Table 11.6, and classification results are presented in Table 11.7, 11.8. Confusion matrices of the open test classification are in Table 11.9.

Set	#Utterances	$\overline{T}_{Utter}$ (s)	σT_{Utter} (s)
Devel	2472	4.10	1.60
Open	1371	4.01	1.50

Table 11.6: Means and standard deviations of utterance durations in devel. set and open test set.

Set	Train	CV	Open
# Utterances	2202	270	1371
UER (%)	9.9 (8.7–11.1)	5.6 (2.8–8.3)	1.6 (0.9–2.3)

Table 11.7: MLP classifier – CFV-based classification; closed/open test, merged male and female utterances, train set + CV set = devel set. Mean values followed by 95% confidence intervals in parentheses. Gender-independent task.

Set	Devel FM	Open FM	Devel DM	Open DM
# Utterances	2472	1371	2472	1371
UER (%)	6.6 (5.6–7.6)	2.5 (1.7–3.3)	8.1 (7.0–9.2)	2.8 (1.9–3.6)

Table 11.8: GMLC classifier – CFV-based classification; closed/open test, merged male and female utterances, FM/DM – full/diagonal covariance matrix. Mean values followed by 95% confidence intervals in parentheses. Gender-independent task.

The best performance in the closed test was displayed by FM–GMLC. In the open test, MLP reached the best accuracy, followed by FM–GMLC and DM–GMLC. The MLP's performance gain on GMLC classifiers was statistically significant while the differences between the FM–GMLC and DM–GMLC performances were insignificant.

Set	MLP		FM–GMLC		DM–GMLC	
	Neutral	LE	Neutral	LE	Neutral	LE
Neutral	591	13	600	4	598	6
LE	9	758	30	737	32	735

Table 11.9: GMLC and MLP classifiers – open test confusion matrices, merged male and female utterances. Mean values followed by 95% confidence intervals in parentheses. Gender-independent task.

Consistently for all three systems, the classification accuracy on the open test set reached higher values than on the development set. This somewhat surprising observation can be explained based on the results of the digit recognition experiments conducted in Sec. 9.2.3. There, the development/open sets were subsets of the development/open sets used in the actual classification task. As shown in Table 9.1, the open LE test set WERs were consistently higher than the development LE set WERs. Note that in that particular experiment the 'development' set was neither used for training the recognizer, nor for adjusting the front-ends, and, hence, represented only another open data for the recognizer. This suggests that the open LE set recordings capture a higher level of LE compared to the development LE set. This hypothesis is supported by the results of the actual classification task, where the open LE set utterances are more easily distinguishable from the open neutral utterances compared to the development LE set.

As shown in Table 11.9, false rejections and acceptances[10] are best balanced in the MLP. In the GMLC, further adjustment of the likelihood ratio threshold λ would be required to approach the matrix symmetry.

11.2 TSR Experiment

Finally, a TSR system consisting of the neutral/LE classifier in combination with the neutral/LE-dedicated recognizers was formulated (see Fig. 11.5). Front-ends for the dedicated recognizers were chosen based on the performance tests conducted in Sec. 9.4.2 (see Table 9.5). RFCC–LPC was chosen as a front-end of the LE recognizer due to its superior performance on Lombard speech. For neutral speech, PLP–DCT, PLP, and MFCC–LPC provided the best results. Since the performances were statistically identical in all three cases, the standard PLP features were selected as a front-end for the neutral recognizer[11]. Three TSR variants, employing MLP, FM–GMLC, and DM–GMLC, were compared. The LE-robust RFCC–LPC features were designed for female speech, hence, the TSR

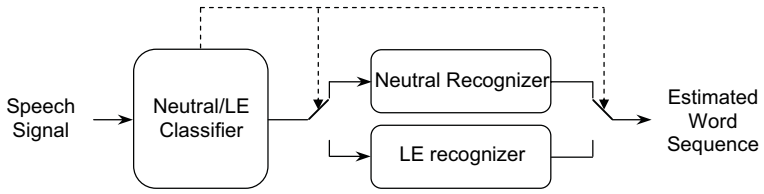

Figure 11.5: Two-Stage Recognition System.

experiments were conducted exclusively on female utterances. Gender dependent acoustic models of the dedicated recognizers were trained on the train set described in Sec. 9.1. The neutral and LE

[10] See Sec. 11.1.1.

[11] It is noted that PLP and RFCC–LPC share, besides the difference in the filter banks used, a similar extraction scheme, which is convenient when optimizing the system implementation.

open test sets were also preserved as in Sec. 9.1, except the files where *SNR Tool*[12] failed to extract *SNR* were excluded, reducing the neutral set from 1450 to 1439 digits and the LE set from 1880 to 1837 digits. The TSR task was applied on the merged neutral and LE open test sets.

To allow for comparison of the TSR performance with the optimal case when all neutral data were processed by the neutral recognizer and all LE data by the LE recognizer, transcriptions obtained from the TSR system were split into two groups, neutral and LE, according to the real talking style origin (neutral/LE) of the input speech files. Results of the TSR experiment are shown in Table 11.10. The first two rows of the table represent the cases when the neutral and LE open test sets were processed by a single recognizer employing either PLP or RFCC–LPC features. The results displayed by PLP on the neutral set and by RFCC–LPC on the LE set represent optimal performances given the systems available in this experiment and given the utterance talking style (neutral/LE) assignments provided by CLSD'05.

Set		Real – neutral	Real – LE
# Female digits		1439	1837
WER (%)	PLP	4.3 (3.3–5.4)	48.1 (45.8–50.4)
	RFCC–LPC	6.5 (5.2–7.7)	28.3 (26.2–30.4)
	MLP TSR	4.2 (3.2–5.3)	28.4 (26.4–30.5)
	FM–GMLC TSR	4.4 (3.3–5.4)	28.4 (26.4–30.5)
	DM–GMLC TSR	4.4 (3.3–5.4)	28.4 (26.3–30.4)

Table 11.10: Performance of single recognizers and TSR systems. Mean values followed by 95% confidence intervals in parentheses.

It can be seen that all three TSR systems reached optimal performance on both neutral and LE sets, significantly gaining on the single recognizers when exposed to mismatched talking styles (PLP/LE speech, RFCC–LPC/neutral speech). MLP TSR displayed an even slightly higher accuracy on the neutral set than the 'optimal' PLP recognizer. This suggests that some utterances labeled in CLSD'05 as neutral or LE were acoustically closer to the opposite talking style and the classifier was able to assign them to a more appropriate style class. No performance loss was observed when switching from FM–GMLC TSR to DM–GMLC TSR, hence, computationally less demanding DM–GMLC may be an efficient classifier choice for TSR-based systems.

11.3 Conclusions

Adapting an ASR system towards a target environmental scenario and talking style may yield considerable performance improvements, however if often comes at a cost of weakening the system's resistance to condition drifts. An example of this phenomenon can be found in Chap. 9 where improving the front-end efficiency for the purposes of LE task was proportional to decreasing its performance in the neutral task (see Fig. 9.3). This chapter addressed such a performance tradeoff by combining neutral/LE specific sub-systems into one unit. At the input stage of the system, talking style comprised in the actual utterance is classified and subsequently, the utterance is passed to the corresponding neutral/LE-dedicated recognizer.

[12] See Sec. 4.3.1.

During the system design, features suitable for gender/phonetic content-independent neutral/LE classification were searched first. Based on the analysis of neutral and LE feature distributions, *SNR*, F_0, σF_0, and *spectral slope* were chosen to form CFV. It was found that linear F_0 representation (Hz) provides better neutral/LE discrimination compared to the logarithmic one (cents of semitones). Discriminative properties of spectral slope extracted from various frequency bands were studied, finding a band of 60 Hz–1 kHz to yield superior results. Subsequently, MLP and Gaussian maximum likelihood neutral/LE classifiers were trained using the proposed CFV. In the open tests, MLP displayed the best classification accuracy. GMLC employing full and diagonal covariance matrices reached comparable results. This suggests that, similarly to GMM/HMM ASR systems, in neutral/LE classification diagonal covariance matrices provide efficient substitutes for computationally more complex full covariance matrices.

Finally, a two-stage recognition system comprising neutral/LE classifier and neutral/LE-dedicated recognizers was formulated. The neutral-dedicated and LE-dedicated recognizers employed PLP and RFCC–LPC, respectively (i.e., features that displayed superior performance for the given talking styles in the previous experiments). Both recognizers were trained using neutral speech samples only. The two-stage recognition system yielded an improvement from 6.5 % to 4.2 % WER on neutral speech and from 48.1 % to 28.4 % WER on LE speech when compared to the dedicated recognizers exposed to adverse talking style (RFCC–LPC recognizer exposed to neutral speech and PLP recognizer exposed to LE speech). Compared to previous studies using style-dependent subsystems to formulate a style-independent recognizer, the newly proposed TSR requires LE data just for training the neutral/LE classifier while only neutral speech samples are required for training the neutral/LE-dedicated recognizers (it is noted that training neutral/LE classifier requires considerably lower amount of data than training acoustic models of an ASR system).

It can be assumed that performance of the neutral/LE classifier is dependent on the duration of the classified utterance. For very short utterances comprising a limited amount of voiced segments, the classification rate can be expected to drop considerably. Evaluation of the duration–WER dependency of the proposed classifier is a matter of future work.

Chapter 12

Conclusions

The goal of this thesis has been to study the speech production variations introduced by Lombard effect and to design algorithms that increase the resistance of ASR systems to these variations. The study was conducted in the following steps:

- Establishing a framework for speech feature tracking and ASR,
- Acquisition of Lombard speech database,
- Analysis of speech feature distributions in neutral and Lombard speech,
- Evaluation of the impact of Lombard effect on the performance of the ASR systems,
- Proposal of methods for Lombard effect-robust ASR.

The following sections summarize the outcomes of these steps and suggest future research directions in the area of Lombard speech recognition.

12.1 Data Acquisition and Feature Analysis

- *Acquisition of CLSD'05*: To address the problem of the sparse occurrence of LE in available Czech speech corpora, a Czech Lombard speech database CLSD'05 was acquired. The newly proposed recording setup motivated speakers to maintain intelligible communication over simulated background noise. The database comprises recordings of neutral and LE utterances from 26 speakers.
- *Compensation for auditory feedback attenuation*: During the CLSD'05 Lombard speech recording, speakers were provided a simulated noisy background through closed headphones. The headphones caused attenuation of the speakers' auditory feedback. To allow for elimination of this undesirable effect in the future recordings, the attenuation by headphones was measured and a transfer function of the compensation speech feedback determined.
- *Design of a time-domain pitch tracker DTFE*: A time-domain pitch tracker DTFE was proposed and compared to five state-of-the-art pitch trackers on the reference database. DTFE displayed a comparable performance to autocorrelation and cross-correlation algorithms on high SNR audio channels while considerably reducing computational requirements of the other methods.
- *Comparison of selected Czech speech corpora*: Detailed analysis of speech collected in neutral and adverse environments was conducted on two commercial databases and on the newly established CLSD'05 database. While the feature shifts due to LE were mostly negligible in the case of the

commercial databases, in CLSD'05, considerable variations of excitation, vocal tract transfer function, and phoneme durations were found, suggesting that CLSD'05 recordings are strongly affected by LE, and hence, valuable for further thesis experiments.

12.2 Newly Proposed Techniques for LE-Robust ASR

In particular, data-driven features, modified VTLN, formant-driven frequency warping, and two-stage recognition employing neutral/LE classification and neutral/LE-specific recognizers were proposed in this thesis. In addition, equalization of LE using a voice conversion system provided by Siemens Corporate Technology, Munich, and model adaptation to LE using adaptation/recognition framework provided by TUL Liberec were evaluated.

The methods were tested in ASR experiments incorporating various subsets of male/female gender-dependent digit/LVCSR tasks, depending on the aims of the given method. A considerably higher ASR deterioration by LE was observed in females, hence, part of the algorithms suppressing an impact of LE on ASR focused on female speech exclusively. Efficiency of the individual LE-suppression approaches in the female digit recognition task is compared in Fig. 12.1. Due to a relatively high number of algorithm variants proposed throughout the thesis, only setups yielding the best performance are presented for each suppression approach. Blue and yellow bars in the figure represent baseline WERs on neutral and Lombard speech, respectively, red bars denote WER on Lombard speech after the LE-suppression techniques were applied.

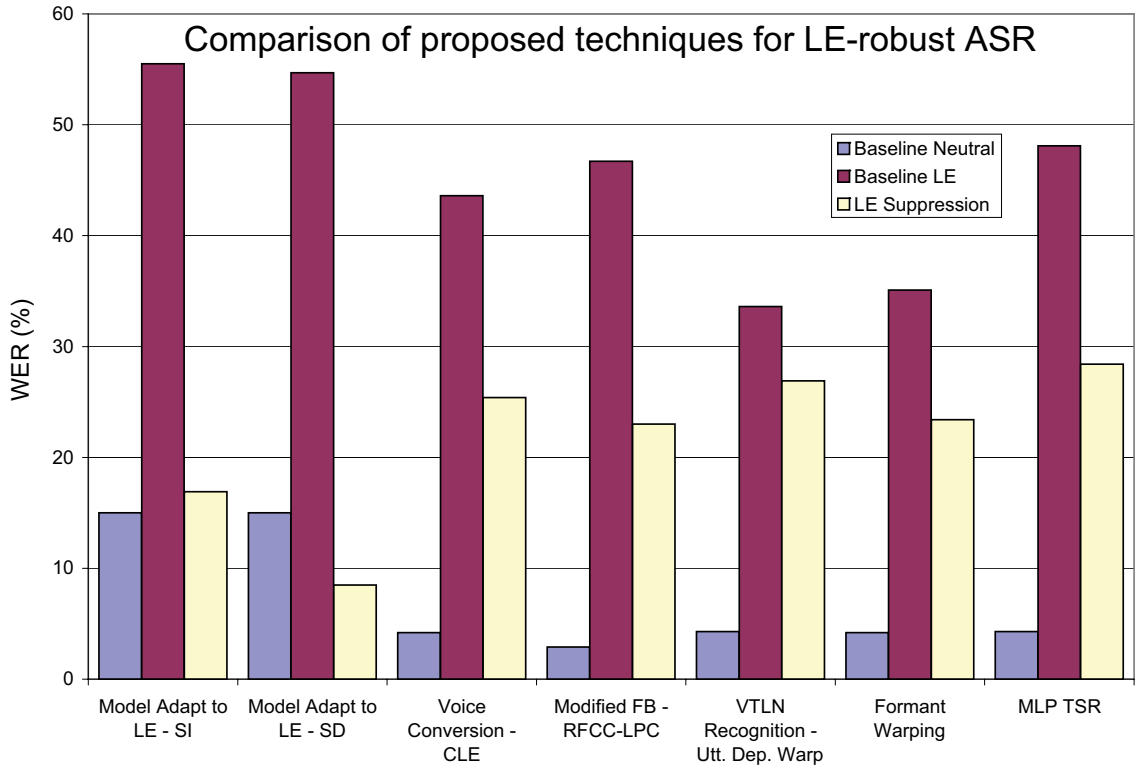

Figure 12.1: A comparison of proposed techniques for LE–robust ASR – female digit recognition task.

Baseline performances varied across the experiments due to differences in the recognizers and data sets used, however, effectiveness of the individual methods can still be easily compared.

- *Acoustic model adaptation* was used to transform means of neutral models towards LE adaptation data[1]. Both speaker-dependent (SD) and speaker-independent (SI) adaptation (Table 7.1 – 'SD adapt to LE' and 'SI adapt to LE – disjunct speakers') provided the highest WER reduction on LE speech compared to the other methods[2]. Model adaptation requires availability of a sufficient amount of labeled adaptation data. In the real-world system, adaptation data can be collected on-the-fly. A presence of strong LE may slow down the adaptation convergence as the accuracy of transcriptions estimated from decoding the adaptation data by the initial (unadapted) neutral acoustic models will be presumably quite low.

- *Voice conversion* (VC) was trained on parallel Lombard/neutral utterances and used for normalizing Lombard speech towards neutral[3]. VC was applied on the speaker-dependent level to transform both formants and excitation component of LE speech towards neutral (CLE) or to transform only excitation component ($CLEF_0$). CLE considerably improved accuracy of female digit recognition (Table 8.6). In the female LVCSR task and in the male digit and LVCSR tasks, CLE was not effective. In some of these cases, $CLEF_0$ helped to reduce WER. The limited success of VC can be attributed to the observed inaccuracies of formant transformation. When applying VC in the real-world ASR, the recognition system will have to contain a speaker identification unit to choose actual speaker-dependent voice conversion models (which will increase the complexity of the system).

- *Features employing modified filter banks* – contribution of frequency sub-bands to speech recognition performance was analyzed, yielding curves of linguistic information distribution across frequency. Based on the information distributions, new front-end filter bank replacements for MFCC and PLP features were designed, providing superior robustness to LE. PLP with modified filter bank (RFCC–LPC) reached the second best performance improvement[4] on LE speech (after model adaptation) when employed in a neutral-trained recognizer, see Fig. 12.1 (Table 9.5). While RFCC–LPC was designed exclusively for female LE speech, it can be expected that the proposed design scheme will be effective also for male speakers. No modifications of an ASR system architecture are needed when replacing a standard front-end by RFCC–LPC. Furthermore, novel 20Bands–LPC features replacing the bark-distributed trapezoidal filter bank in PLP by 20 equidistant rectangular filters were proposed. 20Bands–LPC considerably outperformed MFCC and PLP on Lombard speech while preserving performance of the former features on neutral speech. This suggests that auditory models-based filter banks do not necessarily represent the optimal choice for ASR front-ends.

- *Vocal tract length normalization* (VTLN) – a modified vocal tract length normalization method was proposed. In the standard VTLN recognition procedure, the warping factor is determined with respect to the utterance transcription estimated using the normalized models. However, the mismatch between Lombard speech to be recognized and normalized neutral models would result in very inaccurate utterance transcription estimates and warping estimates. The modified approach addresses this issue by considering all transcription estimates obtained from the

[1]The framework was provided and the model adaptation conducted by Dr. Petr Červa, Technical University of Liberec. Author of the thesis designed the experiments and provided data for adaptation.

[2]In the adaptation task, the baseline neutral and LE WERs are higher than in other experiments because here, a LVCSR system was used for the digit recognition.

[3]Using voice conversion (VC) for normalization of Lombard speech was proposed by Prof. Harald Höge, Siemens Corporate Technology, Munich, Germany. David Sündermann (Siemens) provided the VC system (Sündermann *et al.*, 2006b) and conducted the system training and data conversion. Author of the thesis provided data for VC training and for conversion, analyzed impact of VC on speech features, and evaluated VC efficiency in the ASR experiments.

[4]The baseline WERs shown for RFCC–LPC in Fig. 12.1 refer to the performance of the recognizer with unmodified PLP.

warping grid and picking the most likely one. The modified VTLN provided significant improvements of recognition accuracy on both female and male LE speech (Table 10.1). Performance of utterance-driven and speaker-driven warping was compared. The results shown in Fig. 12.1 refer to the performance of VTLN recognition employing utterance-driven frequency warping. Compared to the standard GMM/HMM ASR systems, a multiple decoding pass for each utterance is required in VTLN recognition. The number of the decoding repetitions is given by the interval in which the optimal warping candidate is searched and by the coarseness of the search grid. The impact of limiting the speech bandwidth on distribution and evolution of frequency warping factors was also studied.

- *Formant-driven frequency warping* – a linear function mapping mean LE F_1–F_4 to the neutral ones was searched and used to warp incoming LE utterances before entering a neutral ASR system. For both females and males, formant-driven warping displayed considerable performance gains (Table 10.6). In the presented experiments, the gender-dependent warping function was determined from a relatively large amount of open test set data. To reach more flexible warping, the initial formant-mapping function can be updated on-the-fly by the estimates obtained from the incoming utterances. Formant-driven warping assumes that a reliable formant tracking is available.

- *Classification of neutral/LE speech*: Based on the speech feature distributions found in neutral and LE speech, a set of gender/lexicon-independent parameters effective for neutral/LE classification was proposed (CFV). Subsequently, multi-layer perceptron-based and Gaussian maximum likelihood-based neutral/LE classifiers were built, reaching utterance classification accuracy on the open test set over 97 %.

- *Two-stage recognition system* (TSR) – adapting an ASR system towards a target environmental scenario and talking style may yield substantial performance gains, however, often at a cost of deteriorating the system's resistance to condition variations. The proposed TSR attempts to exploit the superior performance of the neutral/LE-specific recognizers operating in the matched conditions, while, at the same time, preserving the system's robustness to the neutral/LE condition changes. In the first stage of TSR, neutral/LE classifier decides whether the incoming utterance is neutral or Lombard. Subsequently, the utterance is passed to the corresponding neutral-specific or LE-specific recognizer. This scheme ensures that the neutral/LE-specific recognizers always face utterances of similar or acoustically close talking style. When exposed to the mixture of neutral and LE utterances, TSR significantly outperformed isolated neutral-specific and LE-specific recognizers (Table 11.10). The proposed TSR system requires LE data just for training the neutral/LE classifier while only neutral speech samples are required for training the neutral/LE-dedicated recognizers (note that training neutral/LE classifier requires considerably lower amount of data than training acoustic models of an ASR system).

12.3 Future Directions

Although many of the approaches and algorithms proposed in this thesis extend concepts suggested by previous studies on LE, they still represent initial steps for an ultimate solution to the LE-robust ASR. Future studies might consider to explore, among others, the following topics:

- On-the-fly model adaptation and formant-driven warping: To increase the recognizer's resistance to speaker/talking style changes, acoustic models and front-end warping functions can be adapted to the incoming data on-the-fly. Since the adaptation parameters will be estimated from a limited amount of samples, update rate constrains assuring convergence of the adaptation have to be established.

- Two-stage system employing a codebook of LE level-specific recognizers: The type and level of background noise as well as the communication scenario (distance between a speaker and listeners, number of listeners) affect the rate of speech production changes. Extending the current TSR setup for a set of recognizers addressing separately different levels of LE may bring further performance gains. Based on the experimental results presented in this thesis, using fixed LE level-dependent equalization, LE level-specific front-ends, or model adaptation may be effective means for obtaining the neutral-trained LE level-specific recognizers,

- A comprehensive analysis of speech feature variations in various types and levels of environmental noise: In the past works, continuous dependencies of noise level/vocal effort (Lombard function) and vocal effort/fundamental frequency were found, see Chap. 3. Analyzing how other speech parameters (spectral slope, formants) vary with the level and type of background noise would be valuable for the design of automatic speech normalization driven exclusively by the estimated noise parameters,

- Neutral/LE talking style assessment in spontaneous speech employing lexical information: (Lee and Narayanan, 2005) has shown that in spontaneous speech, speakers tend to use specific 'emotionally' salient words to express emotions. It can be assumed that the presence of strong noisy background may also affect the vocabulary of spontaneous speech. E.g., speakers may prefer using words more intelligible in noise rather than words containing prevailing number of consonants or confusable words. In such a case, lexical information could complement acoustic features as a cue to the neutral/LE style classification.

Bibliography

Abe, M., Nakamura, S., Shikano, K., and Kuwabara, H. (**1988**). "Voice conversion through vector quantization", in *Proc. of ICASSP'88*, 655–658 (New York, USA).

Acero, A. and Droppo, J. (**1998**). "Method and apparatus for pitch tracking", European Patent EP 1 145 224 B1, 22.11.1999.

Ahmadi, S. and Spanias, A. (**1999**). "Cepstrum-based pitch detection using a new statistical V/UV classification algorithm", IEEE Transactions on Speech & Audio Processing **7**, 333–338.

Aldrich, J. (**2006**). *Earliest Known Uses of Some of the Words of Mathematics*, chapter Eigenvalue, eigenfunction, eigenvector, and related terms. Jeff Miller (Ed.), URL `http://members.aol.com/jeff570/e.html`.

Arlinger, S. D. (**1986**). "Sound attenuation of TDH-39 earphones in a diffuse field of narrow-band noise", The Journal of the Acoustical Society of America **79**, 189–191.

Arslan, L. M. (**1999**). "Speaker transformation algorithm using segmental codebooks (STASC)", Speech Communication **28**, 211–226.

Arslan, L. M. and Hansen, J. H. L. (**1997**). "A study of temporal features and frequency characteristics in American English foreign accent", The Journal of the Acoustical Society of America **102**, 28–40.

Baker, J. (**1975**). "The DRAGON system–an overview", IEEE Transactions on Acoustics, Speech, and Signal Processing **23**, 24–29.

Baudoin, G. and Stylianou, Y. (**1996**). "On the transformation of the speech spectrum for voice conversion", in *Proc. of ICSLP'96*, volume 3, 1405–1408 (Philadelphia, Pennsylvania).

Bell, C., Fujlsaki, G., Heinz, J., Stevens, K., and House, A. (**1961**). "Reduction of speech spectra by analysis-by-synthesis techniques", The Journal of the Acoustical Society of America **33**, 1725–1736.

Biem, A. and Katagiri, S. (**1997**). "Cepstrum-based filter-bank design using discriminative feature extraction training at various levels", in *Proc. of ICASSP'97*, volume 2, 1503–1506 (Washington, DC, USA).

Blomberg, M. and Elenius, D. (**2004**). "Comparing speech recognition for adults and children", in *Proc. of Fonetik'04*, 156–159 (Stockholm, Sweden).

Boersma, P. (**1993**). "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound", in *Proc. of the Institute of Phonetic Sciences 17*, 97–110 (University of Amsterdam, Nederlands).

Boersma, P. and Weenink, D. (**2006**). "Praat: Doing phonetics by computer (version 4.4.33)", [Computer program] .

Boll, S. (**1979**). "Suppression of acoustic noise in speech using spectral subtraction", IEEE Transactions on Acoustics, Speech, and Signal Processing **27**, 113–120.

Bond, Z. and Moore, T. (**1990**). "A note on loud and Lombard speech", in *Proc. of ICSLP'90*, 969–972 (Kobe, Japan).

Bond, Z. S., Moore, T. J., and Gable, B. (**1989**). "Acoustic–phonetic characteristics of speech produced in noise and while wearing an oxygen mask", The Journal of the Acoustical Society of America **85**, 907–912.

Bou-Ghazale, S. E. and Hansen, J. (**1998**). "HMM-based stressed speech modeling with application to improved synthesis and recognition of isolated speech under stress", IEEE Transactions on Speech & Audio Processing **6**, 201–216.

Bou-Ghazale, S. E. and Hansen, J. H. L. (**1996**). "Generating stressed speech from neutral speech using a modified CELP vocoder", Speech Communication **20**, 93 – 110.

Bou-Ghazale, S. E. and Hansen, J. H. L. (**2000**). "A comparative study of traditional and newly proposed features for recognition of speech under stress", IEEE Transactions on Speech & Audio Processing **8**, 429–442.

Bořil, H. (**2003**a). "Guitar MIDI converter", Master's thesis, CTU in Prague, In Czech.

Bořil, H. (**2003**b). "Pitch detector for guitar MIDI converter", in *Proc. POSTER 2003*, EI1 (Prague, Czech Republic).

Bořil, H. (**2007**). "Normalization of Lombard effect", Research Report No. R07-2, 52 pages, CTU in Prague & Siemens Corporate Technology, Munich.

Bořil, H., Bořil, T., and Pollák, P. (**2006**a). "Methodology of Lombard speech database acquisition: Experiences with CLSD", in *Proc. of LREC 2006 – 5th Conference on Language Resources and Evaluation*, 1644 – 1647 (Genova, Italy).

Bořil, H. and Fousek, P. (**2007**). "Influence of different speech representations and HMM training strategies on ASR performance", Acta Polytechnica, Journal on Advanced Engineering **46**, 32–35.

Bořil, H., Fousek, P., and Höge, H. (**2007**). "Two-stage system for robust neutral/Lombard speech recognition", in *Proc. of Interspeech'07*, 1074–1077 (Antwerp, Belgium).

Bořil, H., Fousek, P., and Pollák, P. (**2006**b). "Data-driven design of front-end filter bank for Lombard speech recognition", in *Proc. of ICSLP'06*, 381 – 384 (Pittsburgh, Pennsylvania).

Bořil, H., Fousek, P., Sündermann, D., Červa, P., and Žďánský, J. (**2006**c). "Lombard speech recognition: A comparative study", in *Proc. 16th Czech-German Workshop on Speech Processing*, 141–148 (Prague, Czech Republic).

Bořil, H. and Pollák, P. (**2004**). "Direct time domain fundamental frequency estimation of speech in noisy conditions", in *Proc. EUSIPCO 2004*, volume 1, 1003 – 1006 (Vienna, Austria).

Bořil, H. and Pollák, P. (**2005**a). "Comparison of three Czech speech databases from the standpoint of Lombard effect appearance", in *ASIDE 2005, COST278 Final Workshop and ISCA Tutorial and Research Workshop* (Aalborg, Denmark).

Bořil, H. and Pollák, P. (**2005**b). "Design and collection of Czech Lombard Speech Database", in *Proc. of Interspeech'05*, 1577–1580 (Lisboa, Portugal).

Brüel and Kjaer (**2004**). "PULSE X – sound & vibration analyzer", URL `http://www.bksv.com/pdf/bu0228.pdf`.

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

Castellanos, A., Benedi, J. M., and Casacuberta, F. (**1996**). "An analysis of general acoustic-phonetic features for spanish speech produced with the lombard effect", Speech Communication **20**, 23–35.

Charpentier, F. J. and Stella, M. G. (**1986**). "Diphone synthesis using an overlap-add technique for speech waveforms concatenation", in *Proc. of ICASSP'86*, 2015 – 2018 (Tokyo, Japan).

Chen, Y. (**1987**). "Cepstral domain stress compensation for robust speech recogniton", in *Proc. of ICASSP'87*, 717–720 (Texas, Dallas).

Chi, S. and Oh, Y. (**1996**). "Lombard effect compensation and noise suppression for noisy Lombard speech recognition", in *Proc. of ICSLP'96*, volume 4, 2013–2016 (Philadelphia, PA).

Childers, D. G. (**1995**). "Glottal source modeling for voice conversion", Speech Communication **16**, 127–138.

Childers, D. G. and Lee, C. K. (**1991**). "Vocal quality factors: Analysis, synthesis, and perception", The Journal of the Acoustical Society of America **90**, 2394–2410.

Chou, W. and Juan, B. H. (**2003**). *Pattern Recognition in Speech and Language Processing* (CRC Press).

Clarke, C. and Jurafsky, D. (**2006**). "Limitations of MLLR adaptation with Spanish-accented English: An error analysis", in *Proc. of ICSLP'06*, 1117–1120 (Pittsburgh, PA, USA).

Cohen, M. A., Grossberg, S., and Wyse, L. L. (**1995**). "A spectral network model of pitch perception", The Journal of the Acoustical Society of America **98**, 862–879.

Cover, T. and Hart, P. (**1967**). "Nearest neighbor pattern classification", IEEE Transactions on Information Theory **13**, 21–27.

Crosswhite, K. (**2003**). "Spectral tilt as a cue to word stress in Polish, Macedonian and Bulgarian", in *Proc. of XV International Conference of the Phonetic Sciences*, 767–770 (Barcelona, Spain).

Cummings, K. and Clements, M. (**1990**). "Analysis of glottal waveforms across stress styles", in *Proc. of ICASSP'90*, volume 1, 369–372 (Albuquerque, USA).

CZKCC (**2004**). URL `www.temic-sds.com`.

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

Deng, L., Cui, X., Pruvenok, R., Huang, J., Momen, S., Chen, Y., and Alwan, A. (**2006**). "A database of vocal tract resonance trajectories for research in speech processing", in *Proc. of ICASSP'06*, volume 1 (Toulouse, France).

Deng, L., Lee, L., Attias, H., and Acero, A. (**2004**). "A structured speech model with continuous hidden dynamics and prediction-residual training for tracking vocal tract resonances", in *Proc. of ICASSP'04*, volume 1, 557–560 (Montreal, Canada).

Dines, J. (**2003**). "Model based trainable speech synthesis and its applications", Ph.D. thesis, Queensland University of Technology, Queensland.

Dreher, J. J. and O'Neill, J. (**1957**). "Effects of ambient noise on speaker intelligibility for words and phrases", The Journal of the Acoustical Society of America **29**, 1320–1323.

Duifhuis, H., Willems, L. F., and Sluyter, R. J. (**1982**). "Measurement of pitch in speech: An implementation of Goldstein's theory of pitch perception", The Journal of the Acoustical Society of America **71**, 1568–1580.

Dunn, H. K. (**1950**). "The calculation of vowel resonances, and an electrical vocal tract", The Journal of the Acoustical Society of America **22**, 740–753.

Dunn, H. K. (**1961**). "Methods of measuring vowel formant bandwidths", The Journal of the Acoustical Society of America **33**, 1737–1746.

ECESS (**2007**). URL `http://www.ecess.eu`.

Eide, E. and Gish, H. (**1996**). "A parametric approach to vocal tract length normalization", in *Proc. of ICASSP'96*, volume 1, 346–348 (IEEE Computer Society, Los Alamitos, CA, USA).

ELRA (**2008**). "European language resources association: Speecon databases", URL `http://catalog.elra.info/search_result.php?keywords=speecon&language=en&osCsid=66`.

Faneuff, J. and Brown, D. (**2003**). "Noise reduction and increased VAD accuracy using spectral subtraction", in *Proc. of the 2003 IEEE International Signal Processing Conference*, Paper number 213 (Dallas, Texas).

Faria, A. and Gelbart, D. (**2005**). "Efficient pitch-based estimation of VTLN warp factors", in *Proc. of Eurospeech'05*, 213–216 (Lisbon, Portugal).

Flanagan, J. L. (**1956**). "Automatic extraction of formant frequencies from continuous speech", The Journal of the Acoustical Society of America **28**, 110–118.

Flanagan, J. L. (**1957**). "Note on the design of "terminal-analog" speech synthesizers", The Journal of the Acoustical Society of America **29**, 306–310.

Fletcher, H. and Munson, W. A. (**1933**). "Loudness, its definition, measurement and calculation", The Journal of the Acoustical Society of America **5**, 82–108.

Fousek, P. (**2007**). "CTUCopy – universal speech enhancer and feature extractor", URL `http://noel.feld.cvut.cz/speechlab/start.php?page=download&lang=en`.

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

Gales, M., Pye, D., and Woodland, P. (**1996**). "Variance compensation within the MLLR framework for robust speech recognition and speaker adaptation", in *Proc. of ICSLP'96*, volume 3, 1832–1835 (Philadelphia, Pennsylvania).

Gales, M. and Young, S. (**1996**). "Robust continuous speech recognition using parallel model combination", IEEE Transactions on Speech & Audio Processing **4**, 352–359.

Gales, M. J. F. and Woodland, P. (**1996**). "Mean and variance adaptation within the MLLR framework", Computer Speech & Language **10**, 249–264.

Gallardo, A., Macías, J., Ferreiros, J., de Córdoba, R., Montero, J., San-Segundo, R., and Pardo, J. (**2003**). "A comparison of several approaches to the feature extractor design for ASR tasks in telephone environment", in *Proc. of the XVth International Congress of Phonetic Sciences*, 1345–1348 (Barcelona, Spain).

Garau, G., Renals, S., and Hain, T. (**2005**). "Applying vocal tract length normalization to meeting recordings", in *Proc. of Interspeech'05*.

Garnier, M., Bailly, L., Dohen, M., Welby, P., and Loevenbruck, H. (**2006**). "An acoustic and articulatory study of Lombard speech: Global effects on the utterance", in *Proc. of Interspeech'06*, 1–7 (Pittsburgh, Pennsylvania).

Gauvain, J.-L. and Lee, C.-H. (**1994**). "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains", IEEE Transactions on Speech & Audio Processing **2**, 291–298.

Ghitza, O. (**1988**). "Auditory neural feedback as a basis for speech processing", in *Proc. of ICASSP'88*, 91–94 (New York, USA).

Gillick, L. and Cox, S. (**1989**). "Some statistical issues in the comparison of speech recognition algorithms", in *Proc. of ICASSP'89*, volume 1, 532–535 (Glasgow, Scotland).

Gold, B. and Rabiner, L. (**1969**). "Parallel processing techniques for estimating pitch periods of speech in the time domain", The Journal of the Acoustical Society of America **46**, 442–448.

Gramming, P., Sundberg, S., Ternström, S., and Perkins, W. (**1987**). "Relationship between changes in voice pitch and loudness", STL-QPSR **28**, 39–55.

Hansen, J. and Bou-Ghazale, S. (**1997**). "Getting started with SUSAS: A Speech Under Simulated and Actual Stress database", in *Proc. of Eurospeech'97*, volume 4, 1743–1746 (Rhodes, Greece).

Hansen, J. and Bria, O. (**1990**). "Lombard effect compensation for robust automatic speech recognition in noise", in *Proc. of ICSLP'90*, 1125–1128 (Kobe, Japan).

Hansen, J. and Clements, M. (**1989**). "Stress compensation and noise reduction algorithms for robust speech recognition", in *Proc. of ICASSP'89*, 266–269 (Glasgow, Scotland).

Hansen, J. and Womack, B. (**1996**). "Feature analysis and neural network-based classification of speech under stress", IEEE Transactions on Acoustics, Speech, and Signal Processing **4**, 307–313.

Hansen, J. H. L. (**1988**). "Analysis and compensation of stressed and noisy speech with application to robust automatic recognition", Ph.D. thesis, Georgia Institute of Technology, Atlanta, GA.

Hansen, J. H. L. (**1994**). "Morphological constrained feature enhancement with adaptive cepstral compensation (MCE–ACC) for speech recognition in noise and Lombard effect", IEEE Transactions on Speech & Audio Processing **2**, 598–614.

Hansen, J. H. L. (**1996**). "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition", Speech Communication **20**, 151–173.

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

Hanson, B. A. and Applebaum, T. H. (**1990**a). "Features for noise-robust speaker-independent word recognition", in *Proc. of ICSLP'90*, 1117–1120 (Kobe, Japan).

Hanson, B. A. and Applebaum, T. H. (**1990**b). "Robust speaker-independent word recognition using static, dynamic and acceleration features: Experiments with Lombard and noisy speech", in *Proc. of ICASSP'90*, volume 2, 857–860 (Albuquerque, USA).

Harris, F. (**1978**). "On the use of windows for harmonic analysis with the discrete Fourier transform", Proc. of the IEEE **66**, 51–83.

Heinz, J. M. and Stevens, K. N. (**1961**). "On the properties of voiceless fricative consonants", The Journal of the Acoustical Society of America **33**, 589–596.

Hermansky, H. (**1990**). "Perceptual linear predictive (PLP) analysis of speech", The Journal of the Acoustical Society of America **87**, 1738–1752.

Hermansky, H. and Fousek, P. (**2005**). "Multi-resolution RASTA filtering for TANDEM-based ASR", in *Proc. of Eurospeech'05*, 361–364 (Lisbon, Portugal).

Hermes, D. J. (**1988**). "Measurement of pitch by subharmonic summation", The Journal of the Acoustical Society of America **83**, 257–264.

Hirsch, H. G. (**2005**). "FaNT - filtering and noise adding tool", URL `http://dnt.kr.hsnr.de/download.html`.

Hirsch, H. G. and Pearce, D. (**2000**). "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions", in *ISCA ITRW ASR2000 'Automatic Speech Recognition: Challenges for the next Millenium'* (Paris, France).

Hönig, F., Stemmer, G., Hacker, C., and Brugnara, F. (**2005**). "Revising perceptual linear prediction (PLP)", in *Proc. of Eurospeech'05*, 2997–3000 (Lisbon, Portugal).

Iida, A., Iga, S., Higuchi, F., Campbell, N., and Yasumura, M. (**2000**). "A speech synthesis system with emotion for assisting communication", in *Proc. ISCA Workshop on Speech and Emotion*, 167–172 (Belfast, Northern Ireland).

Iskra, D., Grosskopf, B., Marasek, K., van den Huevel, H., Diehl, F., and Kiessling, A. (**2002**). "Speecon - speech databases for consumer devices: Database specification and validation", in *Proc. of LREC'2002* (Las Palmas, Spain).

ITU (**1996**). "ITU – P.58 recommendation - head and torso simulator for telephonometry", URL `http://www.itu.int/rec/T-REC-P.58-199608-I/E`.

Janer, L. (**1995**). "Modulated Gaussian wavelet transform based speech analyser (MGWTSA) pitch detection algorithm (PDA)", in *Proc. of Eurospeech'95*, 401–404 (Madrid, Spain).

Jankowski, C., Hoang-Doan, J., and Lippmann, R. (**1995**). "A comparison of signal processing front ends for automatic word recognition", IEEE Transactions on Speech & Audio Processing **3**, 286–293.

Jones, N. (**1999**). *Survey Networks Lecture Notes, Chap. Error Ellipses* (University of Melbourne, Department of Geomatics), URL `http://www.sli.unimelb.edu.au/nicole/surveynetworks/02a/notes09_01.html`.

Jr., J. R. D., Hansen, J. H. L., and Proakis, J. G. (**2000**). *Discrete-time Processing of Speech Signals* (Macmillan Publishing Company, New York).

Junqua, J. and Anglade, Y. (**1990**). "Acoustic and perceptual studies of Lombard speech: Application to isolated-words automatic speech recognition", in *Proc. of ICASSP'90*, volume 2, 841–844 (Albuquerque, USA).

Junqua, J.-C. (**1993**). "The Lombard reflex and its role on human listeners and automatic speech recognizers", The Journal of the Acoustical Society of America **93**, 510–524.

Junqua, J.-C. (**2002**). *Sources of Variability and Distortion in the Communication Process. Robust Speech Recognition in Embedded Systems and PC Applications*, 1–36 (Springer, Netherlands).

Junqua, J.-C., Fincke, S., and Field, K. (**1998**). "Influence of the speaking style and the noise spectral tilt on the Lombard reflex and automatic speech recognition", in *Proc. of ICSLP'98* (Sydney, Australia).

Junqua, J.-C., Fincke, S., and Field, K. (**1999**). "The Lombard effect: A reflex to better communicate with others in noise", in *Proc. of ICASSP'99*, volume 4, 2083–2086 (IEEE Computer Society, Los Alamitos, CA, USA).

Jurafsky, D. and Martin, J. H. (**2000**). *Speech and Language Processing – An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition* (Prentice Hall, Englewood Cliffs, New Jersey).

Kain, A. and Macon, M. (**1998**). "Spectral Voice Conversion for Text-to-Speech Synthesis", in *Proc. of ICASSP'98*, 285–288 (Seattle, USA).

Kay, S. (**1979**). "The effects of noise on the autoregressive spectral estimator", IEEE Transactions on Acoustics, Speech, and Signal Processing **27**, 478– 485.

Kent, R. D. and Read, C. (**1992**). *The Acoustic Analysis of Speech* (Whurr Publishers, San Diego).

Kinnunen, T. (**2002**). "Designing a speaker-discriminative filter bank for speaker recognition", in *Proc. of ICSLP'02*, 2325–2328 (Denver, USA).

Klatt, D. H. (**1977**). "Review of the ARPA speech understanding project", The Journal of the Acoustical Society of America **62**, 1345–1366.

Kleijn, W. B., Bäckström, T., and Alku, P. (**2003**). "On line spectral frequencies", IEEE Sig. Proc. Letters **10**, 75–77.

Kopec, G. (**1986**). "Formant tracking using hidden Markov models and vector quantization", IEEE Transactions on Acoustics, Speech, and Signal Processing **34**, 709–729.

Korn, T. S. (**1954**). "Effect of psychological feedback on conversational noise reduction in rooms", The Journal of the Acoustical Society of America **26**, 793–794.

Kotnik, B., Höge, H., and Kačič, Z. (**2006**). "Manually pitch-marked reference PMA, PDA database based on SPEECON Spanish", Design File v1.0, SIEMENS AG, University of Maribor.

Kwon, O.-W., Chan, K., Hao, J., and Lee, T.-W. (**2003**). "Emotion recognition by speech signals", in *Proc. of Eurospeech'03*, volume 1, 125 – 128 (Geneva, Switzerland).

Lane, H. and Tranel, B. (**1971**). "The Lombard sign and the role of hearing in speech", J. of Speech and Hearing Research **14**, 677–709.

Lane, H., Tranel, B., and Sisson, C. (**1970**). "Regulation of voice communication by sensory dynamics", The Journal of the Acoustical Society of America **47**, 618–624.

Lane, H. L., Catania, A. C., and Stevens, S. S. (**1961**). "Voice level: Autophonic scale, perceived loudness, and effects of sidetone", The Journal of the Acoustical Society of America **33**, 160–167.

LDC (**2008**). "Linguistic data consortium (ldc): Distribution of the speech under simulated and actual stress (susas) database", URL `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC99S78`.

Lee, C. M. and Narayanan, S. S. (**2005**). "Toward detecting emotions in spoken dialogs", IEEE Transactions on Speech & Audio Processing **13**, 293–303.

Lee, L. and Rose, R. (**1996**). "Speaker normalization using efficient frequency warping procedures", in *Proc. of ICASSP'96*, volume 1, 353–356 (IEEE Computer Society, Los Alamitos, CA, USA).

Lei, X., Siu, M., Hwang, M.-Y., Ostendorf, M., and Lee, T. (**2006**). "Improved tone modeling for Mandarin broadcast news speech recognition", in *Proc. of ICSLP'06*, 1237–1240 (Pittsburgh, PA, USA).

Lippmann, R. P., Martin, E. A., and Paul, D. B. (**1987**). "Multi-style training for robust isolated-word speech recognition", in *Proc. of ICASSP'87*, 705–708 (Texas, Dallas).

Liu, D.-J. and Lin, C.-T. (**2001**). "Fundamental frequency estimation based on the joint time-frequency analysis of harmonic spectral structure", IEEE Transactions on Speech & Audio Processing **9**, 609–621.

Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M., and Harper, M. (**2006**). "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies", IEEE Transactions on Speech & Audio Processing **14**, 1526–1540.

Lockwood, P. and Boudy, J. (**1991**). "Experiments with a non-linear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars", in *Proc. of Eurospeech'91*, 79–82 (Genova, Italy).

Lombard, E. (**1911**). "Le signe de l'elevation de la voix", Ann. Malad. Oreille, Larynx, Nez, Pharynx **37**, 101–119.

Mak, B., Tam, Y.-C., and Li, P. (**2004**). "Discriminative auditory-based features for robust speech recognition", IEEE Transactions on Speech & Audio Processing **12**, 27–36.

Markel, J. (**1972**). "Digital inverse filtering-a new tool for formant trajectory estimation", IEEE Transactions on Audio and Electroacoustics **20**, 129– 137.

McCandless, S. (**1974**). "An algorithm for automatic formant extraction using linear prediction spectra", IEEE Transactions on Acoustics, Speech, and Signal Processing **22**, 135– 141.

Mehaffey, J. (**2007**). "Error measures", URL `http://gpsinformation.net/main/errors.htm`.

Mokbel, C. E. and Chollet, G. (**1995**). "Automatic word recognition in cars", IEEE Transactions on Speech & Audio Processing **3**, 346–356.

Monsen, R. B. and Engebretson, A. M. (**1977**). "Study of variations in the male and female glottal wave", The Journal of the Acoustical Society of America **62**, 981–993.

Murray, I. R. and Arnott, J. L. (**1993**). "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion", The Journal of the Acoustical Society of America **93**, 1097–1108.

Neiberg, D., Elenius, K., Karlsson, I., and Laskowski, K. (**2006**). "Emotion Recognition in Spontaneous Speech Using GMMs", in *Proc. of ICSLP'06*, 809–812 (Pittsburgh, PA, USA).

Norskog, L. (**2007**). "Sound exchange tool manual", URL `http://sox.sourceforge.net`.

Nouza, J., Ždánský, J., David, P., Červa, P., Kolorenc, J., and Nejedlová, J. (**2005**). "Fully automated system for Czech spoken broadcast transcription with very large (300k+) lexicon", in *Proc. of Interspeech'05*, 1681–1684 (Lisbon, Portugal).

Novotný, J. (**2002**). "Trénování a využití kontextově závislých HMM modelů fonémů (training and application of context-dependent HMM phoneme models)", Research Report, CTU in Prague, Prague, Czech Republic.

Olsen, P. A. and Dharanipragada, S. (**2003**). "An efficient integrated gender detection scheme and time mediated averaging of gender dependent acoustic models", in *Proc. of Eurospeech'03*, 2509–2512 (Geneva, Switzerland).

Papoulis, A. (**2001**). *Probability, Random Variables and Stochastic Processes* (McGraw-Hill, New York).

Patel, R. and Schell, K. W. (**2008**). "The influence of linguistic content on the Lombard effect", Journal of Speech, Language, and Hearing Research **51**, 209–220.

Paul, D. B. (**1987**). "A speaker-stress resistant HMM isolated word recognizer", in *Proc. of ICASSP'87*, 713–716 (Texas, Dallas).

Pick, H. L., Siegel, G. M., Fox, P. W., Garber, S. R., and Kearney, J. K. (**1989**). "Inhibiting the Lombard effect", The Journal of the Acoustical Society of America **85**, 894–900.

Pickett, J. M. (**1956**). "Effects of vocal force on the intelligibility of speech sounds", The Journal of the Acoustical Society of America **28**, 902–905.

Pinson, E. N. (**1963**). "Pitch-synchronous time-domain estimation of formant frequencies and bandwidths", The Journal of the Acoustical Society of America **35**, 1264 – 1273.

Pisoni, D., Bernacki, R., Nusbaum, H., and Yuchtman, M. (**1985**). "Some acoustic-phonetic correlates of speech produced in noise", in *Proc. of ICASSP'85*, volume 10, 1581– 1584 (Tampa, Florida).

Pittman, A. L. and Wiley, T. L. (**2001**). "Recognition of speech produced in noise", Journal of Speech, Language, and Hearing Research **44**, 487–496.

Pollák, P. (**2002**). "Efficient and reliable measurement and evaluation of noisy speech background", in *Proc. EUSIPCO 2002*, volume 1, 491–494 (Toulouse, France).

Pollák, P., Vopička, J., and Sovka, P. (**1999**). "Czech language database of car speech and environmental noise", in *Proc. of Eurospeech'99*, 2263–2266 (Budapest, Hungary).

Proakis, J. G. and Manolakis, D. K. (**1995**). *Digital Signal Processing: Principles, Algorithms and Applications* (Prentice Hall, 3rd edition).

Psutka, J., Müller, L., and Psutka, J. V. (**2001**). "The influence of a filter shape in telephone-based recognition module using PLP parameterization", in *TSD'01: Proc. of the 4th International Conference on Text, Speech and Dialogue*, 222–228 (Springer-Verlag, London, UK).

Pye, D. and Woodland, P. C. (**1997**). "Experiments in speaker normalisation and adaptation for large vocabulary speech recognition", in *Proc. of ICASSP'97*, 1047–1050 (Munich, Germany).

QuickNet (**2007**). URL `http://www.icsi.berkeley.edu/Speech/qn.html`.

Rabiner, L. and Juang, B.-H. (**1993**). *Fundamentals of speech recognition* (Prentice-Hall, Inc., Upper Saddle River, NJ, USA).

Rabiner, L. and Schafer, R. (**1978**). *Digital Processing of Speech Signals* (Englewood Cliffs: Prentice Hall).

Rajasekaran, P., Doddington, G., and Picone, J. (**1986**). "Recognition of speech under stress and in noise", in *Proc. of ICASSP'86*, volume 11, 733– 736 (Tokyo, Japan).

Sakoe, H. and Chiba, S. (**1978**). "Dynamic programming algorithm optimization for spoken word recognition", IEEE Transactions on Acoustics, Speech, and Signal Processing **26**, 43–49.

Schafer, R. W. and Rabiner, L. R. (**1970**). "System for automatic formant analysis of voiced speech", The Journal of the Acoustical Society of America **47**, 634–648.

Schulman, R. (**1985**). "Dynamic and perceptual constraints of loud speech", The Journal of the Acoustical Society of America **78**, S37–S37.

Secrest, B. G. and Doddington, G. R. (**1983**). "An integrated pitch tracking algorithm for speech synthesis", in *Proc. of ICASSP'83*, 1352–1355 (Boston, Massachusetts).

Seneff, S. (**1986**). "A computational model for the peripheral auditory system: Application of speech recognition research", in *Proc. of ICASSP'86*, volume 11, 1983–1986 (Tokyo, Japan).

Shimamura, T. and Kobayashi, H. (**2001**). "Weighted autocorrelation for pitch extraction of noisy speech", IEEE Transactions on Speech & Audio Processing **9**, 727–730.

Sjolander, K. and Beskow, J. (**2000**). "WaveSurfer - an open source speech tool", in *Proc. of ICSLP'00*, volume 4, 464–467 (Beijing, China).

Skowronski, M. D. and Harris, J. G. (**2004**). "Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition", The Journal of the Acoustical Society of America **116**, 1774–1780.

Sluijter, A. M. C. and van Heuven, V. J. (**1996**). "Spectral balance as an acoustic correlate of linguistic stress", The Journal of the Acoustical Society of America **100**, 2471–2485.

Smith, J. O. (**2006**). *Introduction to Digital Filters, August 2006 Edition* (Stanford University, Center for Computer Research in Music and Acoustics (CCRMA)), URL `http://ccrma.stanford.edu/~jos/filters/Relating_Pole_Radius_Bandwidth.html#sec:resbw`.

Sovka, P. and Pollák, P. (**2001**). *Vybrané metody číslicového zpracování signálů (Selected algorithms for digital signal processing)* (CTU Publishing House, Prague).

Stanton, B. J., Jamieson, L. H., and Allen, G. D. (**1988**). "Acoustic-phonetic analysis of loud and Lombard speech in simulated cockpit conditions", in *Proc. of ICASSP'88*, 331–334 (New York, USA).

Stanton, B. J., Jamieson, L. H., and Allen, G. D. (**1989**). "Robust recognition of loud and Lombard speech in the fighter cockpit environment", in *Proc. of ICASSP'89*, 675–678 (Glasgow, Scotland).

Steeneken, H. J. M. and Houtgast, T. (**1980**). "A physical method for measuring speech-transmission quality", The Journal of the Acoustical Society of America **67**, 318–326.

Steeneken, H. J. M. and Houtgast, T. (**1999**). "Mutual dependence of the octave-band weights in predicting speech intelligibility", Speech Communication **28**, 109–123.

Steeneken, H. J. M. and Verhave, J. A. (**2004**). "Digitally controlled active noise reduction with integrated speech communication", Archives of Acoustics **29**, 397–410.

Stephens, L. J. (**1998**). *Theory and Problems of Beginning Statistics* (McGraw-Hill, New York).

Stergiou, C. and Siganos, D. (**1996**). "Neural networks and their uses", Surprise 96 **4**, URL `http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html`.

Sulter, A. M. and Wit, H. P. (**1996**). "Glottal volume velocity waveform characteristics in subjects with and without vocal training, related to gender, sound intensity, fundamental frequency, and age", The Journal of the Acoustical Society of America **100**, 3360–3373.

Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., and Stokes, M. A. (**1988**). "Effects of noise on speech production: Acoustic and perceptual analyses", The Journal of the Acoustical Society of America **84**, 917–928.

Sündermann, D., Bonafonte, A., Ney, H., and Höge, H. (**2005**a). "A Study on Residual Prediction Techniques for Voice Conversion", in *Proc. of ICASSP'05*, 13–16 (Philadelphia, USA).

Sündermann, D., Höge, H., Bonafonte, A., Ney, H., Black, A., and Narayanan, S. (**2006**a). "Text-independent voice conversion based on unit selection", in *Proc. of ICASSP'06*, volume 1, 81–84 (Toulouse, France).

Sündermann, D., Höge, H., Bonafonte, A., Ney, H., and Black, A. W. (**2005**b). "Residual prediction based on unit selection", in *Proc. of ASRU'05, 9th IEEE Automatic Speech Recognition and Understanding Workshop*, 369–374 (San Juan, Puerto Rico).

Sündermann, D., Höge, H., Bonafonte, A., Ney, H., and Hirschberg, J. (**2006**b). "TC-Star: Cross-language voice conversion revisited", in *Proc. of the TC-Star Workshop 2006* (Barcelona, Spain).

Sündermann, D., Strecha, G., Bonafonte, A., Höge, H., and Ney, H. (**2005**c). "Evaluation of VTLN-based voice conversion for embedded speech synthesis", in *Proc. of Eurospeech'05*, 2593–2596 (Lisbon, Portugal).

Suzuki, J., Kadokawa, Y., and Nakata, K. (**1963**). "Formant-frequency extraction by the method of moment calculations", The Journal of the Acoustical Society of America **35**, 1345–1353.

Suzuki, T., Nakajima, K., and Abe, Y. (**1994**). "Isolated word recognition using models for acoustic phonetic variability by Lombard effect", in *Proc. of ICSLP'94*, 999–1002 (Yokohama, Japan).

Takizawa, Y. and Hamada, M. (**1990**). "Lombard speech recognition by formant-frequency-shifted LPC cepstrum", in *Proc. of ICSLP'90*, 293–296 (Kobe, Japan).

Talkin, D. (**1987**). "Speech formant trajectory estimation using dynamic programming with modulated transition costs", The Journal of the Acoustical Society of America **82**, S55–S55.

Talkin, D. (**1995**). *Speech Coding and Synthesis*, chapter A Robust Algorithm for Pitch Tracking (RAPT). W.B. Kleijn and K.K. Paliwal (Eds.), 495–518 (Elsevier, Amsterdam, Netherlands).

Teager, H. M. (**1980**). "Some observations on oral air flow during phonation", IEEE Transactions on Acoustics, Speech, and Signal Processing **28**, 599–601.

Terhardt, E. (**1974**). "Pitch, consonance, and harmony", The Journal of the Acoustical Society of America **55**, 1061–1069.

Terhardt, E., Stoll, G., and Seewann, M. (**1982**). "Algorithm for extraction of pitch and pitch salience from complex tonal signals", The Journal of the Acoustical Society of America **71**, 679–688.

Tian, B., Sun, M., Sclabassi, R. J., and Yi, K. (**2003**). "A unified compensation approach for speech recognition in severely adverse environment", in *Proc. of Uncertainty Modeling and Analysis, ISUMA'03*, 256–261 (College Park, Maryland).

Titze, I. R. and Sundberg, J. (**1992**). "Vocal intensity in speakers and singers", The Journal of the Acoustical Society of America **91**, 2936–2946.

Tukey, J. W. (**1974**). "Nonlinear methods for smoothing data", in *Proc. IEEE Electronics and Aerospace Systems Convention, EASCON'74*, 673 (Washington, DC, USA).

Varadarajan, V. and Hansen, J. H. (**2006**). "Analysis of Lombard effect under different types and levels of noise with application to in-set speaker ID systems", in *Proc. of Interspeech'06*, 937–940 (Pittsburgh, Pennsylvania).

Varga, A. P. and Moore, R. K. (**1990**). "Hidden Markov model decomposition of speech and noise", in *Proc. of ICASSP'90*, volume 21, 845–848 (Albuquerque, New Mexico).

Volkmann, J., Stevens, S. S., and Newman, E. B. (**1937**). "A scale for the measurement of the psychological magnitude pitch", The Journal of the Acoustical Society of America **8**, 208–208.

Vondrášek, M. (**2007**). "SNR Tool", URL `http://noel.feld.cvut.cz/speechlab/cz/download/snr.pdf`.

Vondrášek, M. and Pollák, P. (**2005**). "Methods for speech SNR estimation: Evaluation tool and analysis of VAD dependency", Radioengineering **14**, 6–11.

Vorperian, H. K., Kent, R. D., Lindstrom, M. J., Kalina, C. M., Gentry, L. R., and Yandell, B. S. (**2005**). "Development of vocal tract length during early childhood: A magnetic resonance imaging study", The Journal of the Acoustical Society of America **117**, 338–350.

Wakao, A., Takeda, K., and Itakura, F. (**1996**). "Variability of Lombard effects under different noise conditions", in *Proc. of ICSLP'96*, volume 4, 2009–2012 (Philadelphia, PA).

Webster, J. C. and Klumpp, R. G. (**1962**). "Effects of ambient noise and nearby talkers on a face-to-face communication task", The Journal of the Acoustical Society of America **34**, 936–941.

Wikipedia (**2007**). "Wikipedia", URL `http://en.wikipedia.org`.

Womack, B. and Hansen, J. (**1995**a). "Stress independent robust HMM speech recognition using neural network stress classification", in *Proc. of Eurospeech'95*, 1999–2002 (Madrid, Spain).

Womack, B. and Hansen, J. (**1995**b). "Stressed speech classification with application to robust speech recognition", in *NATO-ESCA Proc. International Tutorial & Research Workshop on Speech Under Stress*, 41–44 (Lisbon, Portugal).

Womack, B. and Hansen, J. (**1996**a). "Classification of speech under stress using target driven features", Speech Communication, Special Issue on Speech Under Stress **20**, 131–150.

Womack, B. and Hansen, J. (**1999**). "N-channel hidden Markov models for combined stress speech classification and recognition", IEEE Transactions on Speech & Audio Processing **7**, 668–677.

Womack, B. D. and Hansen, J. H. L. (**1996**b). "Improved speech recognition via speaker stress directed classification", in *Proc. of ICASSP'96*, volume 1, 53–56 (Atlanta, Georgia).

Xia, K. and Espy-Wilson, C. (**2000**). "A new strategy of formant tracking based on dynamic programming", in *Proc. of ICSLP'00*, volume 3, 55–58 (Beijing, China).

Xu, D., Fancourt, C., and C.Wang (**1996**). "Multi-channel HMM", in *Proc. of ICASSP'96*, volume 2, 841–844 (Atlanta, Georgia).

Yao, K., Paliwal, K. K., and Nakamura, S. (**2004**). "Noise adaptive speech recognition based on sequential noise parameter estimation", Speech Communication **42**, 5–23.

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

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. (**2000**). *The HTK Book Version 3.0* (Cambridge University, Cambridge, England).

Zhou, G., Hansen, J., and Kaiser, J. (**1998**). "Linear and nonlinear speech feature analysis for stress classification", in *Proc. of ICSLP'98*, volume 3, 883–886 (Sydney, Australia).

Zwicker, E. (**1961**). "Subdivision of the audible frequency range into critical bands (Frequenzgruppen)", The Journal of the Acoustical Society of America **33**, 248–248.

Čmejla, R. and Sovka, P. (**2002**). "Introduction to bayesian data classification", Akustické listy, in Czech **8**, 3–10.