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

Applying Articulatory Features within Speech Recognition

Disertation thesis

Ing. Petr Mizera

Ph.D. programme: Electrical Engineering and Information Technology
Branch of study: Electrical Engineering Theory
Supervisor: Doc. Ing. Petr Pollák, CSc.

Prague, August 2019

Thesis Supervisor:

Doc. Ing. Petr Pollák, CSc.
Department of Circuit Theory
Faculty of Electrical Engineering
Czech Technical University in Prague
Technická 2
160 00 Prague 6
Czech Republic

Declaration

I hereby declare I have written this doctoral thesis independently and quoted all the sources of information used in accordance with methodological instructions on ethical principles for writing an academic thesis. Moreover, I state that this thesis has neither been submitted nor accepted for any other degree.

In Prague, August 2019

....
Ing. Petr Mizera

Abstract

This thesis deals with possible applications of Articulatory Features (AF) in speech recognition systems with special focus on improvement of Czech spontaneous speech recognition. As spontaneous speech is caused by frequent occurrence of coarticulation process, assimilation and reduction of phones and as AF contain the information about speech production mechanisms, they might represent a possible way how to improve results of these systems. So, the potential contribution of AF-Based TANDEM ASR architecture on the tasks of the recognition or phonetic segmentation of spontaneous speech is described in this work as well as their performance under more adverse acoustic conditions.

As the first result, the multi-valued AF classes for Czech and four East-European languages (Slovak, Polish, Hungarian, and Russian) were defined and unified and further work was focused on the estimation of AF using artificial neural networks. The suitability of standard and advanced acoustic speech features were analyzed for the AF estimation, mainly from the point of view of temporal context at the input of ANN/DNN network. The optimum length of $210 \div 310$ ms was found across languages. The Czech AF classes were estimated with the average *FAcc* around 90%. The behaviour of AF estimation in mismatched or adverse noisy acoustic conditions was also studied and the robustness of DCT-TRAP features was proved as the best choice for this task.

The application of AF within ASR was realized in the form of AF-Based TANDEM system, however, baseline ASR systems had to be prepared, mainly Czech casual speech recognition system with focus on optimization of acoustic and language models as well as the usage of different corpora resources for this task. The performance of the AF-Based TANDEM system was then analyzed for the English phone recognition and Czech ASR tasks. Positive impact of this system was observed for standard monophone (*mono*) and triphone (*tri1*) systems, which are based on MFCC features. The ASR combination of GMM-HMM/DNN-HMM with the AF-Based TANDEM system on the level of lattice with decoded hypotheses significantly improved baseline results.

Finally, phonetic segmentation task was analyzed using various type of acoustic model architectures (GMM-HMM, DNN-HMM, and AF-based TANDEM) as well as focusing on proper pronunciation variant selection. It was done for the following two task: read English (TIMIT) and casual Czech (NCCCz) and two-stage forced-alignment with combination of DNN-HMM and optimized monophone-based system was proposed and the improvement of phone boundary determination was proved for both tasks. The 93% phone boundaries accuracy on the level of 30ms criteria was achieved for read speech in TIMIT, the accuracy around 90% was achieved on for casual one in NCCCz.

Keywords: articulatory features, automatic speech recognition, casual speech recognition, phonetic segmentation, Kaldi

Abstrakt

Předložená disertační práce se zabývá možnými aplikacemi artikulačních příznaků (AF) v úloze rozpoznávání řeči s užším zaměřením na zlepšení rozpoznávání spontánní a neformální řeči pro češtinu. Zejména řeč vytvořená při neformálních rozhovorech je velmi ovlivněna řadou fonetických jevů jako jsou koartikulace, asimilace či redukce hlásek díky méně přesné artikulaci. Artikulační příznaky, které obsahují informaci o produkci řeči, se proto nabízejí jako jedno z možných řešení pro zlepšení přesnosti rozpoznávání neformální řeči. Práce popisuje potenciální přínos TANDEM systémů založených na AF v úlohách rozpoznávání a fonetické segmentaci spontánních promluv.

Práce se nejprve zabývá definicí a popisem artikulačních příznaků pro český jazyk a čtyři východoevropské jazyky (slovenštinu, polštinu, maďarštinu a ruštinu), pro které byly artikulační třídy sjednoceny. Druhou významou částí práce je implementace klasifikátorů AF tříd z řečového signálu založených na bázi hlubokých neuronových sítích, včetně výběru vhodných akustických příznaků pro odhad AF. Hlavní pozornost je věnována optimálnímu nastavení časového kontextu na vstupu DNN sítě. Výsledkem experimentální části pak je nalezení časového kontextu v rozmezí 210 ÷ 310 ms pro všechny analyzované jazyky. AF třídy byly pro češtinu odhadnuty s přesností 90% na úrovni klasifikace v krátkodobých časových rámcích. Dále byla analýzována přesnost odhadu AF v šumu a nepřizpůsobených akustických podmínkách. Zde se prokázala robustnost DCT-TRAP příznaků a jejich vhodnost pro AF klasifikaci.

V další části práce jsou použity AF v rozpoznávačích na bázi TANDEM architektury s cílovým zaměřením na implementaci a optimalizaci rozpoznávání neformální řeči. Nejprve však byly vytvořeny základní systémy resp. akustické modely (GMM-HMM, DNN-HMM resp. TANDEM). Přesnost AF-TANDEM systému byla studována na rozpoznávání anglických hlásek a české řeči. Experimenty ukázaly pozitivní přínos AF-TANDEM na úrovni monofonních a trifónových systémů. Kombinace GMM-HMM/DNN-HMM a AF-TANDEM systémů s AF-TANDEM systémem ukázala významné zlepšení oproti základnímu systému trénovaného bez AF příznaků na úrovni dekódovaných hypotéz.

Poslední část práce je věnována fonetické segmentaci realizované pomocí různých akustických modelů (GMM-HMM, DNN-HMM a AF-TANDEM) s přihlédnutím na vhodný výběr výslovnostních variant. Experimenty byly provedeny pro dvě úlohy: anglickou čtenou řeč a českou neformální řeč. Zvýšení uspěšnosti automatické segmentace bylo dosaženo algoritmem dvoufázové fonetické segmentace, kde komplexní DNN-HMM systém je použit pro získání fonetického přepisu pro následné automatické zarovnání pomocí optimalizovaného monofonního systému. V případě složitější úlohy, jako je neformální řeč, kombinace dvou typů DNN-HMM systémů vedla na 90% přesnost určení hranic hlásek. Pro anglickou čtenou řeč byly hranice určeny s 93% přesností.

Klíčová slova: Artikulační příznaky, automatické rozpoznávání řeči, rozpoznávání neformální řeči, fonetická segmentace, Kaldi

Acknowledgements

First of all, I would like to thank my supervisor Doc. Ing. Petr Pollák, CSc. for his excellent guidance and support during my Master and PhD studies at the Czech Technical University in Prague. He introduced me to digital speech signal processing and automatic speech recognition areas.

Special thanks belong to my colleague Michal Borský for infinite and fruitful discussions about ASR during our PhD study and for being my true friend. I would also like to thank Jan Bartošek for valuable discussions about Phonetic segmentation and Prosody.

Finally, the special thanks go to my wife Misa for her love and support during this hard journey. Thank you my family for supporting me during my studies at universities.

List of Tables

List of Figures

List of Acronyms

AF Articulatory Features.

ANN Artificial Neural Network.

ASR Automatic Speech Recognition.

CMVN Cepstral Mean Variance Normalization.

DCT Discrete Cosine.

DFT Discrete Fourier Transform.

DNN Deep Neural Network.

FAcc Frame Accuracy.

GMM Gaussian Mixture Models.

HMM Hidden Markov Model.

LDA Linear Discriminant Analysis.

LM Language Model.

LSTM Long Short-Term Memory.

LVCSR Large Vocabulary Continuous Speech Recognition.

MFCC Mel-Frequency Cepstral Coefficients.

ML Maximum Likelihood.

MLLR Maximum Likelihood Linear Regression.

MLP Multilayer Perceptron.

MMI Maximum Mutual Information.

PCA Principal Component Analysis.

PER Phone Error Rate.

PLP Perceptual Linear Prediction.

RNN Recurrent Neural Network.

TRAP Temporal Pattern.

TTS Text To Speech.

VAD Voice Activity Detection.

VTLN Vocal Tract Length Normalization.

WER Word Error Rate.

WFST Weighted Finite State Transducers.

Contents

Chapter 1

Introduction

As the voice represents the most effective way of communication between humans, the idea of developing a system which would allow to process human-machine communication has been a natural interest for researchers from the early 20th century. It will be soon 100 years since a real application first used voice iteration [22]. The huge effort and enthusiasm of speech researchers allowed to transform the dreams about a natural human-machine communication to a real voice technology system which are now used in real life conditions. Of course, the big progress in information technologies and the digital revolution boosted the development of voice technologies. High performance computing clusters containing graphical processing units make it possible to train and deploy voice-driven systems based on Artificial Intelligence (AI).

Various applications are commonly used in our daily lives. The list includes dictation systems [84] enabling us to replace keyboard input with a natural speech, virtual agents and voice-controlled devices to smart homes, cars, or mobile phones [15], on-line subtitling, archiving, or monitoring of broadcast or TV programs [138], analysis of speech in medical application for diagnostic purposes, general identification of speaker, or many others.

The research in the field of automatic speech recognition (ASR) started using speech recognition based on Dynamic Time Warping (DTW) [125]. Since then, the field has made a great progress towards the Large Vocabulary Continuous Speech Recognition (LVCSR) based on hidden Markov models (HMM) which were proposed by Jelinek from IBM laboratories [54] in 1980s. This statistical approach using combined Gaussian Mixture Models (GMM) and HMM has become the state-of-the-art for many years. In 1990s, artificial neural network (ANN) was suggested to replace GMM in HMM modelling. The hybrid ANN/HMM ASR system was proposed by Morgan [91] where a multi-layer perceptron (MLP) network was used to estimate the HMM state-posterior probabilities. This approach was designed to overcome limits of the GMM/HMM approach, specifically the fact that modeled features must have a Gaussian distribution [124]. A TANDEM architecture was later proposed by Hermansky, where the an MLP was used as a classifier generating phoneme posterior probabilities which were then transformed, decorrelated and used as features on the input of standard GMM/HMM-based ASR system [47].

The final great leap, in terms of recognition accuracy, came with the adoption of the deep neural network (DNN) and deep-learning techniques to model the acoustic and the language components of the speech. The Microsoft Research team[21], [48], [23] proposed a system based on context-dependent deep neural network - hidden Markov models (CD-DNN-HMM) [21], [48], [23] and the architecture was, for a time, a mainstream solution for all LVCSR tasks running in real-time. The most recent trend, enabled by an increase in computational power and the availability of huge amount of data, includes ASR that model simultaneously the whole recognition chain, i.e. so called End-To-End ASR systems (proposed by Google in 2015 [16]). End-to-End, and DNN-based ASR systems in general, are currently at the top of research interest and they typically achieve a very high accuracy in normal conditions for a majority of world languages.

The precision of an ASR is worse in the case of spontaneous or casual speech, where the pronunciation of particular words can be strongly reduced. Creating a robust ASR system for spontaneous speech has been a challenge for several decades which makes it a popular research topic. Also, the situation of strongly adverse environmental conditions, i.e. when background noise in recorded speech is very high, can decrease the accuracy of an ASR. Many times, it is also not possible to use too complex architectures (e.g. for embedded systems) or a system development is constrained by a limited amount of available data. In such cases, traditional approaches based on GMM-HMM architecture still play a very important role as a looking for an optimization of speech features computed in the front-end part of general ASR system.

The idea to use Articulatory Features (AF) for extension of standard speech features used in ASR is one possible way of how to improve the robustness of spontaneous or noisy speech recognition. The speech production knowledge is a natural part of an ASR system and the examples include using decision trees for triphone state tying or Vocal Tract Length Normalization (VTLN) to reduce speaker variability. It has been showed that including AF for acoustic modelling in a TANDEM system and for pronunciation modelling can achieve very good results in the task of spontaneous speech recognition for the English language [33], [80], [73]. Concerning the Czech language, the issue of spontaneous speech has been worked on for several years [138], [52], [113], [116], [96], [114]. However, the achieved word error rate (WER) is still rather low, about 50%.

Therefore, the general motivation of this work is to find an optimized approach of AF estimation for Czech and other languages and to analyze the contribution of AF in real-life applications, mainly for robust spontaneous and casual speech recognition.

Finally, the thesis is organized as follow.

- Chapter 2 summarizes the state-of-the-art in ASR and discusses usage of AF within speech technology and goals of this thesis are defined.
- Chapter 3 describes the ASR framework and speech databases which are used in the experimental part for training and evaluation of AF classifiers, AM, phone recognizer, and automatic phonetic segmentation. The chapter presents the baseline LCVSR results under various conditions. Finally, the design of ASR system for Czech casual speech recognition task is presented and analyzed.
- Chapter 4 describes the multi-values AF classes for Czech language. The unification of AF for four East-European languages (Slovak, Polish, Hungarian, and Russian) is also presented. The chapter provides a design of AF classifiers with focus on optimum structure of MLP/DNN, type of input features and studies the optimum temporal context at the input of MLP/DNN networks. The accuracy of AF classifiers for all languages is presented.
- Chapter 5 presents the results of experiments focusing on the contribution of AF in the task of phoneme recognition for English and the task of ASR for recognition of Czech read and casual speech. The incorporation of AF information into ASR is analyzed in the form of AF-TANDEM system and using combination of ASR decoded hypotheses. Then, the possible contribution of AF in Clinical applications is discussed.
- Chapter 6 deals with automatic phonetic segmentation of English read speech and Czech casual speech. The chapter analyzes the impact of various types of AMs (GMM-HMM and DNN-HMM) on the accuracy of phone boundaries determination for both languages. The two-stage forced-alignment is proposed, described and analyzed in this chapter. Finally, the chapter describes the process of canonical NCCCz lexicon review and updated the LexFix tool for these purposes. The impact of irregular pronunciation on automatic phonetic segmentation of Nijmegen Corpus of Casual Czech with created lexicon is analyzed.
- Chapter 7 summarizes the content and contributions of this thesis and discusses the next potential work with AF.

Chapter 2

ASR and Articulatory Features: State-of-the-Art

The automatic speech recognition task represents the complex problem which requires knowledge from various research areas such as mathematics, digital signal processing, artificial intelligence, acoustics, phonetics, phonology and linguistics. Therefore, the goal of this chapter is to introduce an ASR system, describe the state-of-the-art ASR architectures, and discuss challenges and motivation for the research described within this thesis.

2.1 Conventional GMM-HMM ASR system

Generally, the goal of the recognizer is to decode an input speech signal which contains an encrypted linguistic message spoken by a speaker to a word sequence. The structure of a modern ASR recognizer consists of five basic modules: feature extraction, a acoustic model, a language model, a pronunciation dictionary, and a decoder, and it is shown in 2.1.

The module of feature extraction segments a discrete sequence of speech samples to short frames and transforms frames from time to frequency or cepstral domain. The feature extraction process removes redundant information from speech signal (fundamental frequency, a phase) and keeps only the information which is important for acoustic modelling. The benefit of cepstral features is that they are decorrelated, which is preferred by GMM-based acoustic modelling, but also due to the fact that similarly sounding speech units create clusters in the cepstral domain. To improve the robustness of cepstral features, cepstral-mean normalization is commonly applied to reduce channel or speaker variably.

The acoustic model describes the acoustic variability of modelled speech units. The

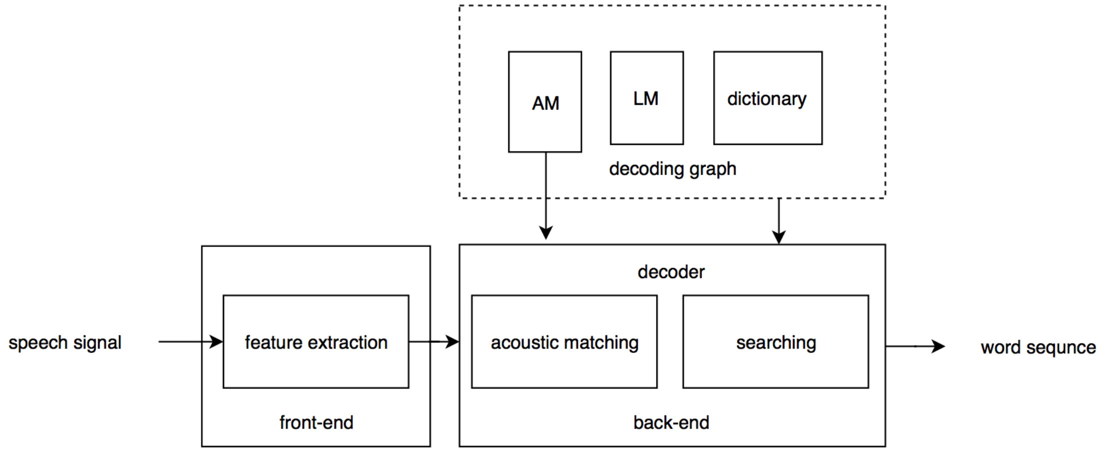


Figure 2.1: The principle structure of the ASR system

selection of the speech units is largely arbitrary but the typical choice is to use monophones (i.e. context-independent phones) or triphones (context-dependent phones). The stochastic approach based on HMM coupled with GMM has been traditionally used to tackle the variability in both time and cepstral domains for several decades.

The purpose of the language model (LM) is to model dependencies between particular words in a sentence. Typically, rule-based or stochastic models are used. The parameters of a stochastic LM are trained on huge text corpora. The dictionary contains mapping between words and theirs pronunciations.

Before the decoding process starts, the acoustic, language models and dictionary are compiled to a complex decoding graph which represents the source of knowledge for the decoding process. The decoding process starts with computation of features which are then passed to a decoder that computes acoustic scores in an acoustic matching block and combines them with language scores that are stored in the recognition graph. The process ends by searching a large hypotheses space for the word sequence which maximizes a posterior probability for given input sequence of features.

From a mathematical point of view, the recognition problem can be formulated using the following equation:

$$\widehat{\mathbf{W}} = \arg\max_{W} P(\mathbf{W}|\mathbf{O}) \tag{2.1}$$

where $\mathbf{W} = w_1, w_2, ..., w_n$ is the sequences of spoken words, $\mathbf{O} = o_1, o_2, ..., o_j$ is the sequences of the feature vectors and $\widehat{\mathbf{W}}$ represents the decoded hypothesis. The conditional probability can be re-written applying Bayes theorem to

$$\widehat{\mathbf{W}} = \arg\max_{W} \frac{P(\mathbf{O}|\mathbf{W})P(\mathbf{W})}{P(\mathbf{O})} \tag{2.2}$$

where the $P(\mathbf{O}|\mathbf{W})$ represents acoustic model, $P(\mathbf{W})$ represents language model and $P(\mathbf{O})$ is the feature vectors probability (it is independent of $\widehat{\mathbf{W}}$).

2.1.1 Cepstral-based speech features

To solve the speech recognition problem means to solve a pattern classification problem. The design of suitable features is critical for this approach. Various methods of feature extraction were designed with regards to understanding how the speech is produced by a human production system and how the speech is perceived by a human auditory system. This chapter describes a category of cepstral features which model spectral properties of a shot-time speech signal. Cepstral coefficients are used as features for speech recognition due to arguments based on speech production and perception knowledge [76]. The first is that the cepstrum deconvolves voice source from the vocal tract. The second is that humans perceive sounds in critical bands. The most common feature set includes the so called mel-frequency or perceptual-linear-predictive cepstral coefficients (MFCC of PLP).

Articulatory features represent another category of ASR features which are based on speech production knowledge. The process of speech production and the AF features is discussed and described in section 2.3.

Short-time spectral analysis

The goal of the feature extraction component shown in 2.1 is to convert speech signal to sequences of features vectors for further classification describing well a variation of non-stationary speech signal characteristics during the time. Discretized and quantized analog speech signal is usually saved in 16bit Pulse-Code Modulation (PCM) format (μ-law/a-law represent formats used in 8k telephone domain). Digital speech signal is then represented as a sequence of samples $x[n] = \{x[0], x[1], ..., x[N-1]\}$, where N is number of samples.

A preemphasis filter is often used as the first step in the extraction process for many speech features. The objective of the preemphasis is to amplify high-frequency components as a compensation for their attenuation during human production [115]. The filter is designed as a 1st order FIR filter described by one coefficient representing the zero point of its transfer function.

The signal is then split into quasi-stationary short-time frames and each short-time frame is multiplied by a weighting window $w[n]$ (usually Hamming window). The typical window length and the shift is 25 ms and 10 ms respectively. The window is applied to compensate for the effect of spectral leakage in the spectrum computed on the basis of the DFT (realized by an FFT, Fast Fourier Transform algorithm). This procedure represents the first part for a majority of feature extraction techniques.

Mel-Frequency Cepstral Coefficients

MFCC cepstral features were designed to model human auditory system. They approximate nonlinear behavior of human perception as a function of frequency and the computation process is shown in Fig. 2.2.

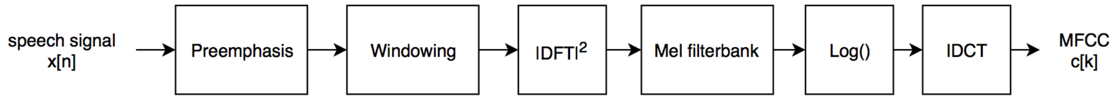


Figure 2.2: MFCC feature extraction

Computed real and imaginary spectral components still contain a lot of redundant information for a classifier, so the phase information is discarded and the magnitude spectrum is transformed using a filter bank to reduce number of spectral components. For MFCC features, this next step involves using the mel-based filter bank which approximates the non-linear perception of frequencies by the human auditory system. Mel-based filter bank consists of overlapping triangular filters which are linearly spread on the mel-frequency scale. These values represent the energy in particular frequency bands. Such a filter bank is usually realized using DFT, where the mel-based logarithmic filter-bank energies g_j are computed as a multiplication of square magnitude spectrum and the frequency response $H_{mel,j}$ of the j-th filter as

$$g_j = ln \sum_{k=0}^{N/2} |S[k]|^2 H_{mel,j}[k] \qquad \text{for } j = 1,\, 2,\, ...,\, M\,, \tag{2.3}$$

where $S[k]$ is DFT-based spectrum, N length of analyzed frame in samples, and M the number of bands in used filter bank. Finally, g_j is a real and even sequence which is transformed to cepstral domain using Discrete Cosine Transform (DCT) to obtain cepstral coefficients c_n, i.e.

$$c_n = \sqrt{\frac{2}{P}} \sum_{j=1}^{P} g_j \cos\left(\frac{\pi n}{P}(j - 0.5)\right) \qquad \text{for } n = 1,\, 2,\, ...,\, M\,. \tag{2.4}$$

The low dimensional MFCC coefficients describe the shape of magnitude spectra which represents the configuration of vocal tract, zeroth-cepstral coefficient c_0 is related to the power of signal, and MFCC coefficients do not contain information about fundamental frequency. The standard practice is to use 13 MFCC for GMM-HMM systems as the application of DCT decorrelates the features on top of transforming them to the cepstral domain. For DNN-HMM system, the filter bank coefficients g_j (FBANK) are sometimes preferable, but a high dimensional MFCC are also used quite often.

Perceptual Linear Prediction

Perceptual Linear Prediction analysis (PLP) represents an alternative to MFCC and it is the second most frequently used approach for speech feature extraction. PLP was developed by Hermansky [46] and the computation process is shown in Fig. 2.3.

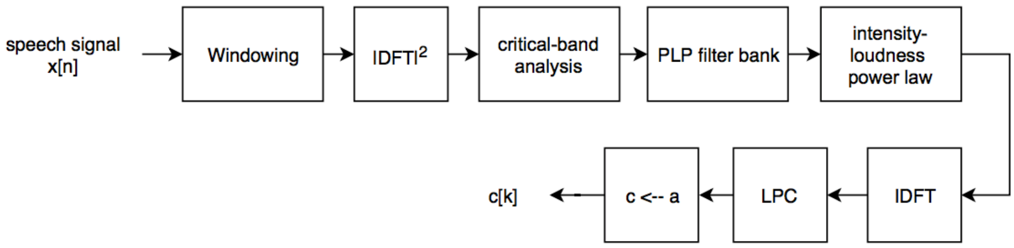


Figure 2.3: PLP feature extraction

The close similarity of MFCC and PLP features can be clearly observed from the diagram. The digitized speech signal is transformed to the short-time power spectral domain and then transformed in several steps which aim to exploit knowledge of the human auditory system (critical-band analysis, equal-loudness-curve and intensity-loudness power-law application). The filter bank is implemented in the power spectral domain and consists of the trapezoidal filters which are equally spaced on the bark scale.

The extraction process ends with the computation of cepstral coefficients using LPC (Linear Predictive Analysis), i.e. modified power spectrum is modeled using all-pole autoregressive (AR) model of the order p. More specifically, inverse DFT transforms power spectra to autocorrelation coefficients and then, Yule-Walker method is used to estimate the autoregressive coefficients a_k. The final PLP cepstral coefficients are obtained from the autoregressive coefficients using recursive formula as

$$c_n = -a_n - \frac{1}{n}\sum_{k=1}^{n-1}(n-k)a_k c_{n-k}, \text{ for } n = 1,\ 2,\ ...,\ p.$$

Temporal context information

It was shown by many authors that the long temporal contextual information is very important for increasing accuracy of ASR systems [92], [103]. The above mentioned cepstral features describe a speech signal in short-time frames which capture temporarily limited information about speech units modelled by a classifier. To capture the longer temporal context and to improve the ASR accuracy, the delta (dynamic), delta-delta (acceleration), or third differential (delta-delta-delta) features are often concatenated into the final feature vector. These features describe the evolution of static features [62] and thus capture the movement of the vocal tract during the phonation. Dynamic features are computed from static features using simple difference formula which is applied across a window

$$d_n = \frac{\sum_{k=1}^{K} k * (c_{n+k} - c_{n-k})}{2 * \sum_{k=1}^{K} k^2}, \tag{2.5}$$

where d_n represents the delta coefficients for current feature frame c_n and K determines the size of the contextual window $(2K+1)$. Its typical value is 2. The acceleration as well as the third differential coefficients are computed in the same manner as the delta coefficients.

Another approach to integrate the long temporal contextual information into ASR that is typically used in TANDEM system [47] is the use of the TempoRAl Pattern (TRAP) based features. The TANDEM system and DCT-TRAP features are described in 2.2.2.

Normalization techniques

Normalization methods are commonly used in ASR or speaker recognition systems to compensate for convolution noise distortion which is caused by channel and speaker variability (e.g. different type of microphone for AM training and testing) or by stationary additive noise. The set of possible options includes two basic techniques. The first option is the Cepstral Mean Normalization (CMN) where mean cepstrum is subtracted from each short-time static cepstral vector. Second option is the Cepstral Mean and Variance Normalization (CMVN) where the subtraction of average cepstrum is followed by a scaling (dividing by standard deviation). Both of these techniques can be computed on per utterance or per speaker basis. In the case of per utterance basis, the mean and standard deviation are estimated from all frames within one utterance. The computed stats are then applied on all frames to provide zero-mean and unit-variance cepstral features for next processing [24]. The former is preferred in online ASR systems whereas the later in offline systems.

LDA-based and speaker-dependent features

An alternative option to dynamic features on how to incorporate temporal context into the feature vector is to stack the static features with a context window of length l, where l is the number of preceding and following frames added to the current frame. It is often used as an input to MLP/DNN classifiers in various tasks such as voice activity detection (VAD), phone recognition or DNN based ASR systems. Concerning convention GMM-HMM system, it can be used as well, however, the dimension of stacked features must be reduced using Linear Discriminant Analysis (LDA) [43] as well as decorrelated using Semi-Tied Covariance (STC) [34] transform. The features processed in such a way are suitable for modelling by HMM with diagonal covariance matrix in a GMM-HMM system [108].

This target feature vector of the size 40 can be further speaker-adapted using feature-space Maximum Likelihood Linear Regression (fMLLR) and these features are used in modern LVCSR systems with GMM-HMM architecture as a standard setup. Fig. 2.4 illustrates the whole processing chain from raw MFCC to final features which are fed into the decoding block.

2.1.2 GMM-HMM based Acoustic Modelling

As it was mentioned before, speech signal is a non-stationary signal. The same speaker, when asked to phonate the same speech unit, can produce signals with a large degree of variability due to the speaking rate, speaking style, quality of the pronunciation or state/age of the speaker. To efficiently model this variability in representation, the stochastic approach was introduced very early on and it still represents an important part of ASR systems working in real-live conditions.

The Gaussian Mixture Models and Hidden Markov Models (GMM-HMM) combination represented the first conventional approach for AM which was widely used by the speech community. The HMM model is finite-state automaton which is characterized by the following parameters:

- non-emitting/ emitting states,
- $b_j(\boldsymbol{o})$ emission probability distributions,

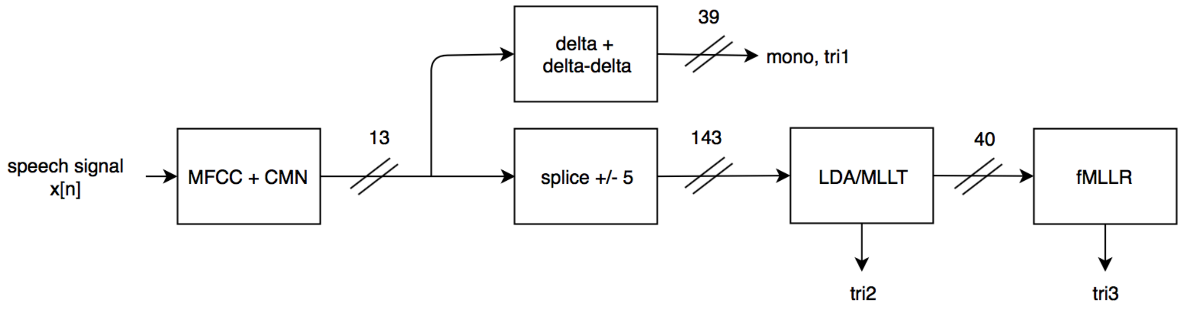


Figure 2.4: Feature extraction used in GMM-HMM systems.

- A transitions matrix,
- initial state probabilities

The structure of a HMM model is defined by a configuration of non-emitting and emitting states. The non-emitting states are used for initial and target states and are mainly for concatenation of particular models to high level structures. The emitting states in HMM are described by a probability density function and they are used to model relationship between HMM state and acoustic observation (e.g. triphones clusters in cepstral domain). The movement between particular states is defined by transitions matrix and its transition probabilities. The commonly used structure for speech unit modelling is a left-right HMM model. The model consists of 3 emitting states without skips which model triphones (8-15 states are used in a word-model) and typically 5 emitting states configuration is used for silence modelling. The GMM distribution was found to be effective for modelling of speech units in cepstral space. An n-dimensional Gaussian function is defined by the following equation

$$\boldsymbol{\mathcal{N}}(\boldsymbol{o},\,\boldsymbol{\mu},\,\boldsymbol{C}) = \frac{1}{\sqrt{(2\pi)^N|\boldsymbol{C}_j|}} \cdot e^{-\frac{1}{2}(\boldsymbol{o}-\boldsymbol{\mu})^T\boldsymbol{C}^{-1}(\boldsymbol{o}-\boldsymbol{\mu})} \tag{2.6}$$

where N is dimension of the Gaussian function which is equal to the dimension of a feature vector $\boldsymbol{o}$, $\boldsymbol{\mu}$ is vector of mean values and $\boldsymbol{C}$ is a covariance matrix which describes shape or rotation of distribution. When particular dimensions in feature vectors are decorrelated, the Gaussian function with diagonal covariance can be used. This is a typical situation, because cepstral features fulfill well above mentioned assumption thanks to DCT in the last step of MFCC computation. Above mentioned Gaussian distribution is used in acoustic model for state description in a generalized form of the GMM, i.e. as the weighted sum of several Gaussian functions (mixtures). Probability density function which describes j-th state of HMM models is then defined as

$$b_j(\boldsymbol{o}) = \sum_{m=1}^{M} c_{jm} \cdot \boldsymbol{\mathcal{N}}\left(\boldsymbol{o},\, \boldsymbol{\mu}_{jm},\, \boldsymbol{C}_{jm}\right) \ , \tag{2.7}$$

where M is number of Gaussian mixture components and c_{jm} is weight of particular Gaussian mixture. To estimate the parameters of a GMM-HMM model, several training algorithms were successfully developed, i.e. Maximum Likelihood (ML) estimation based on Baum-Welch algorithm, discriminative training based on Maximum Mutual Information (MMI) [18], boosted Maximum Mutual Information (bMMI), or Minimal Phone Error (MPE) [55].

2.1.3 Language Modelling & Dictionary

Further important component in ASR is the language model (LM), which covers the linguistic level of spoken and written language and models order of word sequence in recognized utterance. Typical conventional approach used for LM is based on n-grams, which models history of $n1$ words. Regarding the equation 2.2, the goal of the LM is to estimate apriori probability of a word sequence P(W) based on the following equation

$$P(W) = P(w_1)P(w_2|w_1)P(w_3|w_1w_2)....P(w_i|w_1w_2...w_{i-1}) = \prod_{i=1}^{n} P(w_i|w_1,...w_{i-1}) \ , \tag{2.8}$$

where $P(w_i|w_1,...w_{i-1}$ represents the probability of the word i which has dependency on $i-1$ previous words [37].

Since estimating the probability of a word depending on large number of previous words is practically impossible, bigrams or trigrams (i.e 2-grams or 3-grams) represent the most frequently used LMs for decoding within conventional ASR. The higher orders of n-grams (4-grams or 5-grams) are typically used during re-scoring of decoded hypotheses. The n-gram models are trained on large text corpora, but to deal with the problem of sparse data for some n-grams, smoothing techniques must be involved. The commonly used smoothing methods are Good-Turing (GT), Kneser-Ney (KN) and Witten-Bell (WB). Another typical problem of a LM is the mismatch between train corpora and target domain for ASR. Therefore, the LM interpolation methods are used to adapt a generic LM to the target recognition domain.

Additional kind of a special LM is a class-based n-gram model, where the words with similar meaning can be included in a class. Typical examples is a class based LM for ASR designed to control devices, e.g. GPS commands, where we can have class for countries, cities or streets. Nowadays, some systems use LMs based on neural networks where recurrent (RNN) or long short-term memory (LSTM) neural networks are used to model longer dependencies more efficiently [145].

As the purpose of this thesis is the study of an impact of special features within conventional ASR system, the basic setup based on n-gram LMs are discussed in the corresponding experimental parts of this thesis.

The next important component in a ASR system related to language-level description is pronunciation dictionary containing pronunciations for every decoded word. In fact, it joins the modelling at the level of AM and LM respectively. Pronunciation dictionaries contain usually a mapping from orthographic level of a word to canonical phonetic form, i.e. regular pronunciation. In case a lexicon containing good representation of multiple recognition domains is created, automatic methods called g2p (grapheme-to-phoneme) conversion can be used to learn this mapping. For some languages with strong and regular relationship between written and pronounced form of a word, g2p conversion can be defined on an expert level. Czech is such a language and all further described work uses] such g2p tool.

2.1.4 Recognition graph, decoding

The acoustic model, language model, and lexicon represent the key resources of knowledge which need to be put together and saved in an effective format for a searching process in the recognizer. These particular modules of ASR systems are nowadays typically based on Weighted Finite State Transducers (WFST) which allows to simplify the representation of above mentioned parts. A final static recognition graph is then given by a composition of particular automata as

$$HCLG = H \circ C \circ L \circ G \tag{2.9}$$

where $HCLG$ represents final WFST decoding graph, H is HMM topology for modelled triphones, C represents conversion of monophones to context-dependent phones, L is a lexicon mapping words → monophones, and G is a probabilistic grammar or a stochastic language model. The symbol $\circ$ is used for the composition operation which is applied to join above mentioned parts, i.e. automata (modules) H, C, L, and G.

A minimization and determination operation are involved to remove redundant paths from the final decoding graph. This step has significant impact of a size of the graph and the speed of decoding and it represents the main reason to use WFST approach. The process of compilation and minimization of a decoding graph is not a goal of the thesis too and it is used as a standard module in the experimental part. More details can be found in the following resources [90], [109], [107]. Finally, the token-passing algorithm is typically used to find the best hypotheses for a given observation feature vector and the decoded output can be represented using one-best/n-best results or a lattice [109].

2.2 DNN-based speech recognition

The section describes two widely used approaches for integration of artificial neural networks with a HMM-based ASR architecture.

2.2.1 DNN-HMM-based Acoustic Modelling

An alternative approach to GMM for estimating of emitting probabilities within an HMM is based on artificial neural networks and the architecture is called as ANN-HMM hybrid ASR system and it was introduced early by Morgan [91]. In this configuration, MLP-based ANN structure was used for classification of senons which represent states of context-dependent triphones in an AM. Nowadays, these feed-forward neural networks with several hidden layers, called as DNN were found as the most efficient way for AM [48]. DNN-HMM approach significantly improves the accuracy of an ASR when compared to legacy GMM-HMM architectures and it represents an important milestone in the AM development.

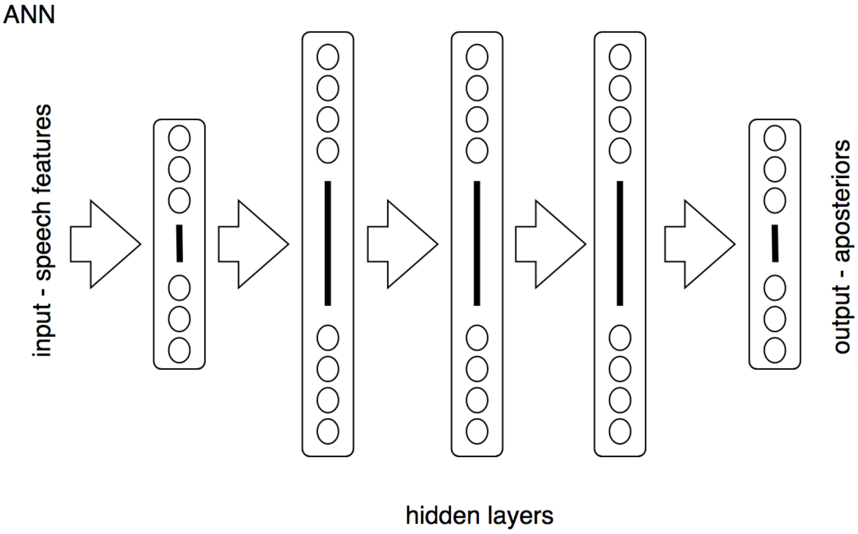


Figure 2.5: Symbolic structure of DNN network.

DNN definition

DNN structure consists of an input layer, several hidden layers and an output layer as shown in Fig. 2.5. The input layer distributes the speech features, and therefore the number of its neurons depends on the size of an input feature vector. Each neuron output in the hidden layer is defined by a commonly used sigmoid activation function

$$f_k(z_k) = \frac{1}{1 + e^{z_k}}, \tag{2.10}$$

which is applied on inner neuron potential z_k computed as a general weighted sum of the neuron inputs

$$z_k = b_k + \sum_{j=1}^{I} w_{jk} x_j, \tag{2.11}$$

where the weights w_{jk} and bias b_k are associated with the k-th neuron. The DNN output represents a posteriori probability of a given HMM state and the possible value is in the range of $0 \div 1$. It is computed by a softmax activation function defined for the k-th output neuron and particular neuron potentials z_j as

$$f_k(z_k) = \frac{e^{z_k}}{\sum\limits_{j=1}^{K} e^{z_j}}. \tag{2.12}$$

The size of the output layer is always given by a number of classified classes (monophones, triphones, senons).

DNN acoustic model and its training

The integration of DNN to HMM based acoustic models is demonstrated in 2.6. As it was mentioned, DNN provides a prediction of context-dependent triphone state probabilities. Computed a posteriori probabilities of the triphone states are also often transformed to likelihood domain by dividing with a priori probabilities of particular triphone states, which is estimated from a train data set.

Concerningthe training of a DNN-HMM acoustic model, it can not be done from scratch as is the case for a GMM-HMM acoustic model. The training process is typically split to two stages. First, the GMM-HMM system is trained to prepare targets for DNN-HMM training obtained by a forced alignment which associates context-dependent triphone states with particular frames. Then, the process of DNN training can start with random initialization of weights or with the initialization of hidden layers based on Restricted Boltzmann Machines (RBMs) [48]. Further, cross-entropy objective function is used together with a backpropagation algorithm. The discriminative training approaches

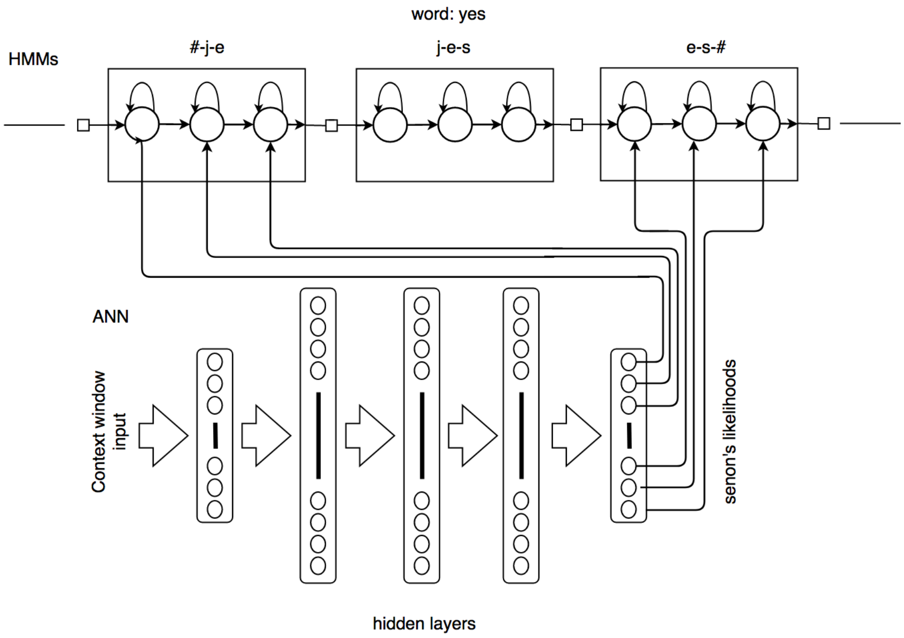


Figure 2.6: Architecture of DNN-HMM ASR system

were adapted for DNN-HMM as well and they represent the next improvement of AM quality [110].

Features used for DNN-HMM architecture

Typically, the short-time feature vectors based on cepstrum and delta and delta-delta features can be used at an input of a DNN. Usually, the features are simply stacked with a temporal context without any further processing as it is not necessary to perform decorrelation for DNN-based models. On the other hand, it is strongly recommended to realize the mean and variance normalization of used features if they are used as an input of a DNN.

Concerning nowadays used DNN-HMM systems, they often use previously mentioned advanced LDA-based features, however, they are still further stacked with the temporal context of several frames to increase the number of input information. An illustrative block scheme of feature extraction procedure using MFCC is in Fig. 2.7.

2.2.2 TANDEM architecture of ASR

Based on the first MLP-HMM based ASR system development, the TANDEM architecture was introduced by Hermansky [47]. The TANDEM system consists of the MLP network, which is used to extend front-end processing and it is followed by the convention GMM-HMM architecture for the AM. Such an architecture is shown in Fig. 2.8. The design

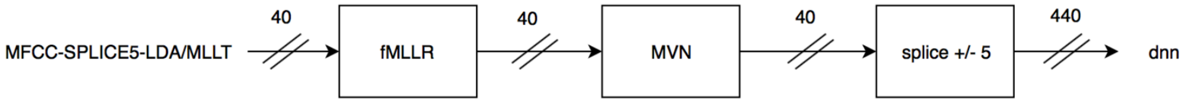


Figure 2.7: Feature extraction used in DNN-HMM systems.

procedure of a TANDEM system can be split into two stages. In the first stage, MLP classifier is trained to map input feature vector to context-independent phoneme classes. Then, the trained MLP classifier is used as a feature extractor component and produces class probabilities per frame. The class probabilities are also called a posterior features. Assuming GMM modelling of features, the logarithm operation and Principal Component Analysis (PCA) are applied on a posterior features to obtain decorrelated features with Gaussian distribution which are suitable for GMM-HMM acoustic modelling. All methods developed for the convention GMM-HMM acoustic modelling (e.g. speaker adaptation or discriminative training) can be used to improve a TANDEM system.

A posterior-feature-based GMM-HMM TANDEM system achieves better results in noisy acoustic conditions when compared to the conventional cepstral-based GMM-HMM or hybrid MLP-HMM ASR systems. However, its performance on ASR tasks in clean conditions were similar to cepstral based GMM-HMM system. Various setups of the feature processing pipeline used within GMM-HMM acoustic modelling were analyzed during last two decades [38], [29], [140]. Simply post-processed a posterior features used within the first TANDEM classifiers are typically combined with standard PLP-Δ-$\Delta\Delta$-HLDA-39 features for LVCSR ASR tasks [7]. Around 2008, the TANDEM system was improved with the bottleneck features (BF), which were proposed by Grezl [38]. The architecture with BF overcame the convention cepstral based GMM-HMM ASR system for the first time in [40], [39]. Later, the more complex hierarchical structure of BF extractors were developed for ASR, speaker recognition or language identification tasks [140], [81], [27].

In the case of features for MLP/DNN classifier, the stacked MFCC/PLP cepstral features or TRAP based features are commonly used to represent longer temporal context. As it was discussed, the purpose of using time context information in ASR is to describe the process of co-articulation produced within human speech production.

TANDEM system was first to apply TRAP feature-extraction technique which is based on using the temporal trajectories of spectral power in the individual critical bands. The technique was proposed by Hermansky and Sharma in 1998 and the analysis of various variants continued in [139], [29]. Both mel or bark filter banks are used and the temporal

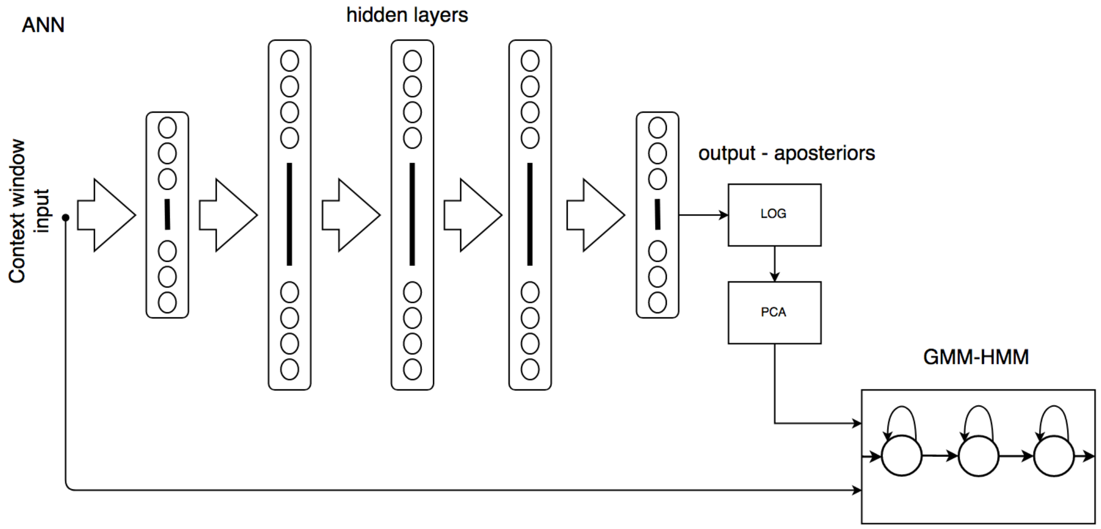


Figure 2.8: Architecture of TANDEM ASR system

trajectories of critical band energies are independently classified or compressed using DCT.

First, the signal is filtered with a preemphasis filter, followed by the auditory spectral analysis. Typically, the number of bands in the auditory spectral analysis is smaller than within the standard MFCC computation and this setup was found to be a good compromise. It enabled to decrease a computational complexity of a MLP and it still proved to yield sufficient accuracy in the classification task. The temporal patterns were created within a context window per each critical band and temporal patterns were compressed using the DCT to decrease the input vector dimension and to increase their decorrelation. In the end, concatenated DCT coefficients represent the DCT-TRAP feature vector on the input of classifier. The significance of contextual information in the task of phone recognition was analyzed in detail for English language in [103], [102], [128] and found to be around 90-110 ms.

As discussed before, the authors in [128], [38] showed that TRAP based features can significantly improve the performance of ASR systems and phone recognition and they have became common for front-end processing in the TANDEM systems. Only one work [112] has applied TRAP to the estimation of AF. In addition, an inclusion of longer context at the input of an ANN-based AF classifier was discussed in [137]. As a result, the next chapter presents the analysis of DCT-TRAP-based estimation of AF for Czech and English language.

2.3 Articulatory Features within Speech technology

Previously discussed approaches of speech representation based on spectral or cepstral analysis respectively are used in many ASR systems with a very good performance under standard conditions. However, when acoustic conditions as well as speaking style with analyzed utterance are worse, these standard features reach their limit due to the possible removal of particular details of a signal characteristic during smoothing procedure as well as due to removing some further information, e.g. about fundamental frequency, etc.

An option on how to improve the representation of a speech signal under more adverse conditions is to include useful information about human speech production or articulation. We are speaking about so called Articulatory Features and they represent one possible way of how speech production knowledge can be incorporated into the ASR systems. Within further sections, speech production process is briefly summarized as well as known applications of AF within speech recognition system are discussed.

2.3.1 Speech Production

The speech signal is an acoustic signal determined by the air flow which comes out of the mouth and which represents the basis of a speech generation. It is generated by the asynchronous movement of various muscles in the voice tract and therefore, the speech signal is a non-stationary signal, however, it can be considered as quasi-stationary signal in short-time periods. The quasi-stationary character allows to process the digitized speech signal using DFT framework in spectral domain as was discussed in section 2.1.1 with regards to auditory features. The quasi-stationary character can be clearly observed in time and spectral domain in Fig. 2.9.

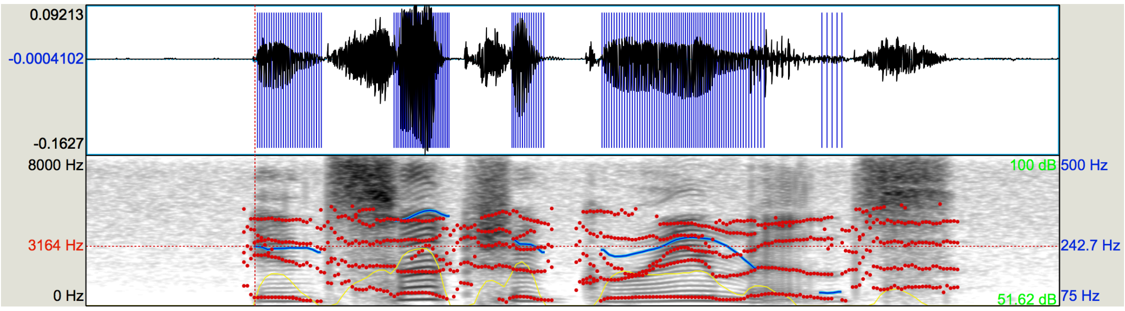


Figure 2.9: Speech signal - time/frequency domain

The whole process of speech production starts in the respiratory system, where the air flow is expelled from lungs, continues to pass through the trachea into the larynx where the vocal tract is located. An important part of the larynx related to speech generation are the vocal folds. They modulate the rising air-flow. When the vocal cords are tight,

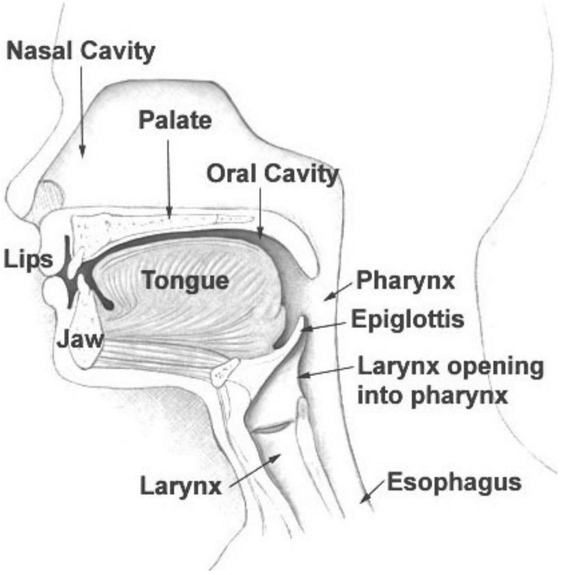


Figure 2.10: Speech vocal tract (source [3])

the rising air-flow generates a glottal cycle which includes glottal open instant and closure instant (pitch marks). The pitch marks can be automatically detected and they are represented using blue lines in Fig 2.9. They are commonly used in the pitch-synchronous segmentation during building of TTS systems or are used for speed modification which is widely used as data augmentation method in an AM [67]. It means that the periodic vibration of the vocal folds creates the voice source signal, which forms the basis of the human voice.

The vibration period of the vocal cords is called the fundamental period or the pitch period T_0 and its inverse value $f_0 = 1/T_0$ is the fundamental frequency. The fundamental frequency is also commonly used in ASR systems for Asian languages and it can be automatically estimated via auto-correlation based methods [36]. The estimated contour of f_0 is shown as the blue line in Fig. 2.9.

The fundamental frequency is an important basic feature of voiced sounds. The voiceless voices are generated by a stream of air passing through the open vocal folds and therefore do not contain the fundamental tone. If the vocal cords and vocal tract are completely calm and open, only breathing occurs. Consequently, the voicing allows to distinguish between voiced and unvoiced speech consonants, so it represents an important AF class.

Concerning further speech production process, the air-flow is modulated in the vocal tract cavities and it is radiated out as acoustic sound waves from the mouth and nostrils. The vocal tract represents a key component of speech production and its anatomy is shown in Fig 2.10. It consists of various organs which are called articulators. The acoustic resonance frequencies of those cavities are also called the formants (the estimated formant frequencies are plotted as red line in spectral domain in Fig. 2.9), which help to distinguish among individual phonemes.

The articulation system consists of oral, nasal and pharyngeal cavities. The pharynx, the soft palate, the hard palate, the alveolar ridge, the tongue, the teeth and the lips represent organs which are involved in the speech production process. The articulation organs have an direct impact on a character of produced speech sound and allow to distinguish among particular vowels and consonants. The tongue is an active organ in the vocal tract and its position has an impact on a creation most of speech sounds. Typically, when the tongue touches the soft palate the velar consonants are created, when alveolar ridge is touching the tongue alveolar consonants are produced or the dental consonants are created when the tongue touches the front teeth [122]. The changes in shape and position of the articulators takes some time, which is referred to as the co-articulation. Position of articulators as well as their possible movement is the bases to define AF classes. Likewise. The vocal tract and movements of articulators is unique for each speaker which allow to use them as human characteristics in Voice Biometry.

Signal model of speech production

To model the speech production process, the pulses/noise generators and a digital filter need to be defined. Digital filter model allows to generate quasi-stationary speech signal $s[n]$ based on periodic updates of its parameters. The block diagram of the speech production model is shown in Fig. 2.11.

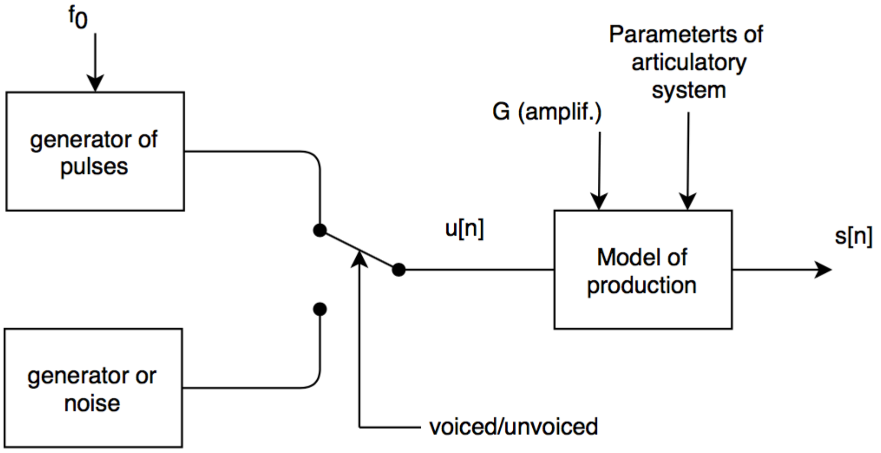


Figure 2.11: Artificial model for generation of speech signal

The pulses (with desired fundamental frequency f_0) or noise generators (white noise) represent the vocal cords and their task is to simulate a voice/unvoiced character of airflow. The shape of target spectrum is modeled by digital filter representing the vocal tract. Typically, an all-pole filter is used for modelling the resonators (cavities) in a articulatory system and its parameters can be estimated using an LPC. The volume is simulated by the parameeter G (amplification coefficient).

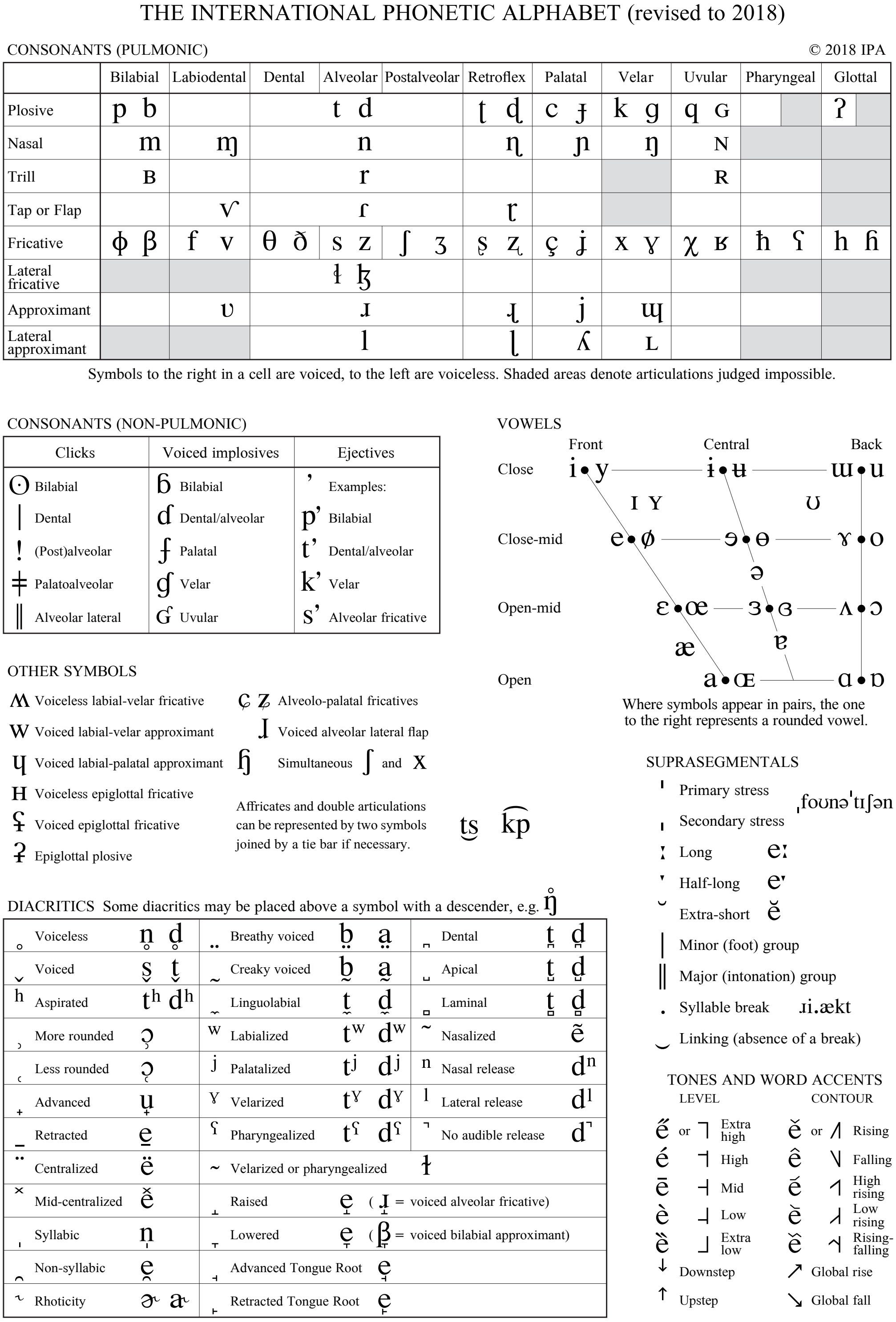
2.3.2 Speech description at phonetic level

The alternative description of the speech production is using the articulatory phonetics and phones. The phone is the acoustic realization of a phoneme, which represents a phonological unit, which allows to distinguish between words. The Czech language has 9 allophones. The phones allow to describe a speech sound with regards to the configuration of the articulation system. Each language has its specific phones set and the unification of language specific phones created of The International Phonetic Alphabet (IPA). The IPA charts is shown in Fig. 2.12. The vowels and consonants classes in IPA table are widely used as articulately features for ASR systems. The difference between standard MFCC/PLP speech features and AFs is that AF describe properties of vocal rather tract than properties of acoustic signal. Therefore, the AF contains complementary information about the speech production.

2.3.3 Applications of Articulatory Features

Articulatory features contain useful information about human speech production or articulators and represent one of the ways how speech production knowledge can be incorporated into the ASR systems. Summary and challenges of using the speech production knowledge in ASR systems are synoptically mentioned by Kirchhoff in [65], Livescu in [78], [80], King and Frankel in [61], Metze [85] and Mitra [87]. The most important arguments for using speech production knowledge in ASR can be summarized in the following points:

- speech data can by used more effectively because some AFs can be shared across a group of phonemes,
- AFs enable better modelling of the co-articulation, assimilation and reduction process which are present especially in spontaneous or casual speech [33],
- AFs can help in adverse conditions when background or convolutional noise is present [65], [66], [86],
- AFs are more suitable for the usage in multilingual tasks in comparison to the usage of language-universal phoneme set.



THE INTERNATIONAL PHONETIC ALPHABET (revised to 2018)

CONSONANTS (PULMONIC)

© 2018 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

Clicks	Voiced implosives	Ejectives
ʘ Bilabial	ɓ Bilabial	ʼ Examples:
ǀ Dental	ɗ Dental/alveolar	pʼ Bilabial
ǃ (Post)alveolar	ʄ Palatal	tʼ Dental/alveolar
ǂ Palatoalveolar	ɠ Velar	kʼ Velar
ǁ Alveolar lateral	ʛ Uvular	sʼ Alveolar fricative

VOWELS

Front Central Back

Close i • y — ɨ • ʉ — ɯ • u

ɪ ʏ ʊ

Close-mid e • ø — ɘ • ɵ — ɤ • o

ə

Open-mid ɛ • œ — ɜ • ɞ — ʌ • ɔ

æ ɐ

Open a • ɶ — ɑ • ɒ

Where symbols appear in pairs, the one to the right represents a rounded vowel.

OTHER SYMBOLS

ʍ Voiceless labial-velar fricative

w Voiced labial-velar approximant

ɥ Voiced labial-palatal approximant

ʜ Voiceless epiglottal fricative

ʢ Voiced epiglottal fricative

ʡ Epiglottal plosive

ɕ ʑ Alveolo-palatal fricatives

ɺ Voiced alveolar lateral flap

ɧ Simultaneous ʃ and x

Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary. t͜s k͡p

DIACRITICS Some diacritics may be placed above a symbol with a descender, e.g. ŋ̊

̥ Voiceless	n̥ d̥	̤ Breathy voiced	b̤ a̤	̪ Dental	t̪ d̪
̬ Voiced	s̬ t̬	̰ Creaky voiced	b̰ a̰	̺ Apical	t̺ d̺
ʰ Aspirated	tʰ dʰ	̼ Linguolabial	t̼ d̼	̻ Laminal	t̻ d̻
̹ More rounded	ɔ̹	ʷ Labialized	tʷ dʷ	̃ Nasalized	ẽ
̜ Less rounded	ɔ̜	ʲ Palatalized	tʲ dʲ	ⁿ Nasal release	dⁿ
̟ Advanced	u̟	ˠ Velarized	tˠ dˠ	ˡ Lateral release	dˡ
̠ Retracted	e̠	ˤ Pharyngealized	tˤ dˤ	̚ No audible release	d̚
̈ Centralized	ë	~ Velarized or pharyngealized	ɫ		
̽ Mid-centralized	e̽	̝ Raised	e̝ (ɹ̝ = voiced alveolar fricative)		
̩ Syllabic	n̩	̞ Lowered	e̞ (β̞ = voiced bilabial approximant)		
̯ Non-syllabic	e̯	̘ Advanced Tongue Root	e̘		
˞ Rhoticity	ɚ a˞	̙ Retracted Tongue Root	e̙		

SUPRASEGMENTALS

ˈ	Primary stress	ˌfoʊnəˈtɪʃən
ˌ	Secondary stress	
ː	Long	eː
ˑ	Half-long	eˑ
̆	Extra-short	ĕ
|	Minor (foot) group	
‖	Major (intonation) group	
.	Syllable break	ɹi.ækt
‿	Linking (absence of a break)	

TONES AND WORD ACCENTS

LEVEL			CONTOUR		
e̋ or	˥	Extra high	ě or	˩˥	Rising
é	˦	High	ê	˥˩	Falling
ē	˧	Mid	e᷄	˦˥	High rising
è	˨	Low	e᷅	˩˨	Low rising
ȅ	˩	Extra low	e᷈	˧˦˧	Rising-falling
ꜜ		Downstep	↗		Global rise
ꜛ		Upstep	↘		Global fall

Figure 2.12: The IPA table - 2018 revision

These advantages and recommendations for incorporating speech production knowledge to ASR have motivated researchers for several decades and a lot of research has been done to develop end-to-end ASR systems based on the speech production knowledge (both observation modelling and pronunciation modelling are based on AFs). Therefore, a lot of research was focused on using AFs in other speech applications such as:

- spontaneous, conversational ASR [33],
- pronunciation modelling [77],
- phone recogniton [120], phonetic segmentation [49]
- spoken term detection [111],
- speaker recognition, speaker verification [130],
- robust speech recognition [41],
- voice conversion [11],
- multilingual or cross-lingual ASR [73], [146],
- application for under-resourced languages [8],

All these areas have been thoroughly studied recently. The above mentioned advantages and applications incorporating AFs represent the motivation for this work. The main focus on the Czech language systems lies in the following tasks:

- spontaneous, conversational ASR,
- robust speech recognition,
- phone recogniton,
- phonetic segmentation segmentation.

2.4 Goals of the thesis

On the basis of the above summary of the start-of-the-art in the given research field, the motivation of this work is to study the properties of articulatory features same as their potential contribution in several tasks of speech recognition, phone recognition and phonetic segmentation, both with the main focus on the processing of the casual and spontaneous speech. The challenge attempt to solve the task of spontaneous speech recognition have been employed by the researchers in the speech community for several decades and it is still a very current topic.

The solution to this problem requires using modern techniques of feature extraction, acoustic modelling and also pronunciation and language modelling. The main motivation of this work should be to incorporate the speech production knowledge using articulatory features to the state-of-the-art ASR systems and to develop a robust spontaneous speech recognition system. Special attention will be paid to the systems working for Czech.

Finally, goals of this thesis can be summarized more precisely in the following points.

- The study of articulatory features contribution for speech recognition and phonetic segmentation.
- In general, to explore modern techniques of feature extraction and acoustic modelling with a special focus on the possible contribution of articulatory features to the description of spontaneous and casual speech (with further applications in the field of informal speech recognition).
- More specifically, perform the optimization of AF estimation and AF-based phone recognition for Czech and English (possibly also for other languages); for this purpose the modern approaches based on DNN.
- To contribute to the processing of speech collected under adverse conditions because informal speech is typically produced in the real environment, often with some level of disturbances (background noise, cross-talks, distant speech, etc.); such approach should be then applied also to other robust speech recognition systems (e.g. in the car environment).
- In the application field, to implement an AF-based phone recognition, a phonetic segmentation, and a AF-TANDEM-based ASR system for the task of spontaneous and informal recognition (focused mainly on Czech, but other languages will be also included).
- Concerning the implementation issues, to use modern toolkits available in the speech research community allowing the realization of above mentioned applications, i.e. KALDI, TNet; the use of our private feature extraction tool CtuCopy will be also extended to include newly designed feature extraction techniques; and the created final implementations will be publicly available.
- Regarding the experimental part, to conduct experiments under various acoustic conditions, i.e. standard read speech (database SPEECON, TIMIT), the speech containing higher level of background noise (car speech data), spontaneous speech (data containing technical lectures), and finally casual speech (Nijmegen Corpus of Casual Czech).
- As a by product of experimental part, special selections of suitable speech data, language models or lexica from available resources completed possibly by additional information according to the requirements of above mentioned experiments are also supposed to be prepared; these data should be then publicly available for the research community when it is allowed by particular license conditions.

Chapter 3

Experimental ASR Framework

As a common ASR framework is used for experiments describing contributions of AF in several different applications in the following parts of this thesis, this section provides a brief overview of the used tools and speech corpora together with some baseline ASR results are presented.

There are several software packages used to conduct ASR experiments: i.e. QuickNet, TNet, CtuCopy, Kaldi. The earliest neural network training was done using the QuickNet package [57]. the TNet software is commonly used for parallel training of the neural networs, using either the multithread data-parallelization for CPU or CUDA parallel computing architecture (GPU). The software was moved to nnet1 in Kaldi around 2012 (developed by Karel Vesely from VUT Brno Group [141]). After the presentation of the Kaldi toolkit at ICASSP 2011 in Prague, Kaldi has become the most popular development toolkit for acoustic modelling that has been under continuous development.

3.1 Kaldi toolkit

The Kaldi toolkit is a modern speech recognition toolkit which supports many the-state-of-the-art training techniques both for GMM-HMM and DNN-HMM ASR architectures. Kaldi is distributed under a non-restrictive Apache licence v2.0 and there exist many recipes which allow the user to build an ASR system on freely available corpora or other commercial corpora provided by the ELRA or LDC. These facts, coupled with the efficiency of its implementation, makes Kaldi the default toolkit not only for beginners or a junior researcher but also to all other members of speech research community.

Kaldi is written by C++ language and it requires various external libraries which allow to work with finite-state transducers, linear algebra, etc. (i.e. OpenFst, ATLAS, BLAS and LAPACK supporting both UNIX and Window systems). A parallelization of Kaldi to cluster computation is also supported using Sun GridEngine software and Slurm. More details on the historical overview can be found at [107].

Kaldi recipes are one of the most important feature of the toolk. They demonstrate an example of how to work with the Kaldi executable tools using standard bash scripts and demonstrate the usage of implemented tools for building various ASR systems (conventional diagonal GMM, sGMM, SAT using fMLLR, MMI, MPE, hybrid and tandem based DNN models, etc.). The Kaldi includes complete recipes for ASR, speaker recognition and language identification supporting more than 40 corpora for various languages: English (WSJ, TIMIT, switchboard, rm, etc.), Danish (sprakbanken), Spanish (fisher_callhome), Egyptian (callhome_egyptian), Arabic (gale_arabic), Mandarin Chinese (gale_mandarin, hkust, thchs30), Swahili (swahili), Japanese (csj), Persian speech (farsdat), as well as Czech (vystadial_cz [70]) or other multilingual corpora (GlobalPhone and Babel). The Kaldi recipes are also used as ASR baseline systems for various challenge tasks such as ASPIRE [4], CHIME [5], REVERB [64]. A major portion of this thesis was done using the toolkit.

Although Kaldi supports all popular feature extraction algorithms, our private tool *CtuCopy* was also used within this thesis. It is an universal feature extractor that contains also speech enhancement techniques developed at our department [30]. Within this work, the *CtuCopy* tool was extended by several funcionalities, e.g. the computation of general derivative of static features, the convolution distortion normalization based on CMS, simple trapper catching various lengths of standard features using the context window, etc. [88], [12].

3.2 Used corpora for training & testing of ASR

All experiments were done with various Czech corpora (e.g. Czech SPEECON database, Czech car speech (CZKCC), NCCCz, CtuTest or CzLecDSP), for English language TIMIT database was used. Concerning other languages, SpeechDat-E (Czech, Polish, Slovak, Hungaria and Russia) corpora were used. These corpora will be described briefly in this section.

3.2.1 Czech SPEECON

Czech SPEECON corpus consists of 550 adult speakers with utterances containing phonetically rich sentences and words as well as some other application commands. Recordings were done in several environments such as office, entertainment, public place or a car, and they were collected by four microphones (channel 0 - headset microphone, channel 1 - close distance, channel 2 - medium distance, channel 3 - far distance). The speech signals were digitized by 16 kHz sampling frequency and saved in raw 16bit PCM format.

All utterances were transcribed at an orthographic level [104], a pronunciation lexicon containing all present word forms is also included. The Czech SPEECON database was used for acoustic modelling, automatic phonetic segmentation, AFs estimation and phone recognition.

SPEECON subset for creating and testing AM

The recordings from the office and the entertainment environments were used for training the basic AM. The training subset consisted of phonetically rich sentences and words, general words and phrases, or digit sequences, and they represented typical clean recording conditions. Utterances which are distorted by strong noise or music, e.g. radio in the background) were found by listening tests and removed from the training subset. The subset for testing of AM quality contained the selection of digit sequences. This test subset was, of course, disjunctive to the training subset. More details are summarized in Table 3.1.

corpora	set	speakers	gender (M/F)	sentences	hours
SPEECON_office	train	160	79/81	43396	38.2
	test	21	8/13	620	1
	dev	20	11/9	589	1
SPEECON_office_ent	train	217	100f+110m	58722	51.7

Table 3.1: Data subsets for creating AM

SPEECON subsets used for AF estimation

With regards to the experiments focused on AF classification, the following two subsets were created: the first one containing rather clean speech signals from a standard office environment (OFFICE subset) and the second one with more noisy utterances from a car environment (CAR subset). Utterances with digits and phonetically rich sentences and words from all available recordings were selected for these subsets. Selected data were divided into non-overlapping training, cross-validation (CV), and test sets. Sizes of these subsets are summarized in more details in Table 3.2.

	OFFICE			CAR		
set		sentences	hours	speakers	sentences	hours
training	101	3450	4.99	48	4042	4.40
cross-val.	17	585	0.88	4	101	0.16
test	77	94	0.16	4	39	0.07

Table 3.2: Data subsets for AF experiments

Since SPEECON database contains only information about the orthographic transcription, the AF targets for MLP/DNN learning were obtained using HMM-based forced-alignment. The produced phone boundaries were then used for mapping of phones to articulatory features. For the test sets, phone boundaries were determined both automatically and manually. The reference manual segmentation of testing data was created by engineers with the knowledge of phonetics and phonology.

Noisy subsets used for AF estimation

The SPEECON corpus contains data from various input channels. These data were used for experiments focused on the robustness of the AF estimation for the same utterances collected with the microphones of different quality. SNR levels strongly vary in these channels, from the average SNR of about 26.82 dB (for channel CS0) representing a clean speech to the average SNR about 6.43 dB (for channel CS3) corresponding to the speech distorted by both convolutional and additive noise. SNR level estimates for all channels are summarized in more details in Table 3.3.

	CS0	CS1	CS2	CS3
OFFICE	26.82	19.51	12.71	6.43
CAR	14.00	6.95	12.06	8.99

Table 3.3: Average values of SNR [dB]

Subsets used for phonetic segmentation

The testing subset for the phonetic segmentation experiments consists of the selection of phonetically rich sentences or digit sequences. The selected utterances represent a rather clean recording conditions from channel CS0. The statistics are summarized in the following Table 3.4.

	Current state			
Sex	minutes	speakers	sentences	phones
Male	1.09	6	16	923
Female	0.20	2	3	172
Total	1.29	8	19	1095

Table 3.4: The evaluation subset statistics

3.2.2 CZKCC - Czech car speech

The CZKCC corpus represents a private database of Czech speech from 1000 speaker recorded in a car environment[1]. The corpus contains car speech recorded in 2 channels

[1] The corpus was collected for TEMIC Speech Dialogue Systems GmbH and Harman Becker Automotive Systems respectively at Czech Technical University in Prague in co-operation with Brno University of Technology and University of West Bohemia in Pilsen.

under various driving conditions using three different microphone setups. In our experiments we use the sub-part containing the speech recorded by a headset microphone only.

The recorded utterances contain different application commands supposed to be used in a car environment as well as phonetic reach sentences. As all phonetically rich material was recorded always in a quiet car (i.e. standing with an engine turn off), these data can be used for the training of general acoustic models. The summary about data subset used in our experiments is in Table 3.5.

corpora	set	speakers	gender (M/F)	sentences	hours
CZKCC_headset	train	244	115/129	10379	16.9
	test	30	14/16	581	1.1
	dev	27	13/14	499	1

Table 3.5: Data subsets for creating AM

3.2.3 NCCCz - Nijmegen Corpus of Casual Czech

The data from the Nijmegen Corpus of Casual Czech (NCCCz) were used for experiments related to spontaneous speech recognition and automatic phonetic segmentation respectively. It contains more than 30 hours of high-quality recordings of casual conversations among 10 triplets of male and 10 triplets of female friends. Sasual speech is defined as a way of talking used within a conversation among close people. All speakers were recorded simultaneously on separate audio channels using cardioid microphones avoiding possible cross-talks in particular channels for each speaker. The speakers were engaged in conversations for approximately 90 minutes and the recordings were obtained by the the procedure described below, which resulted in very informal spontaneous speech data which was presented in [25].

One speaker from each triplet always acted as a confederate who asked two friends of the same gender (henceforth the naive speakers) to participate in recordings of natural conversations. Each session was recorded in a soundproof booth and in the first part of the recording, the confederate pretended to have received an important phone call that had to be answered immediately and the two naive speakers were left alone without information about whether they were already being recorded. Depending on the liveliness of the conversations between the two naive speakers, the confederate returned to the booth.

The second part of the recording consisted of free conversation among the three speakers. Various topics including school, relationships, common hobbies, and stories about all sorts of encounters were addressed. In the third part of the recordings, the experimenter used a list of questions on political and social issues and the speakers were asked to discuss at least four issues from the list and negotiate a common opinion for each question.

This recording procedure of NCCCz was the same as the one used for the collection of similar Dutch, French, or Spanish corpora [135]. The whole corpus has been annotated at orthographic level using standard non-reduced transcription joined by additional marks for non-speech events.

Data subsets for experiments focusing on ASR

corpora	set	speakers	gender (M/F)	sentences	hours
	train	40	20/20	18192	16.1
NCCCz	test	20	10/10	863	1.1

Table 3.6: Data subsets for creating AM

Data subsets for experiments focusing on Phonetic segmentation

The experiments focused on phonetic segmentation were done on utterances from speakers with a standard level of reduced pronunciation and without disturbances such as additive noise, non-speech acoustic events or overlapping speech. These were manually segmented at the phonetic level. This segmentation was created by specialists with knowledge in phonetics and phonology. Consequently, the evaluation subset containing selected utterances from 8 speakers was created. The amount of data in this evaluation subset is summarized in Table 3.7 (values for the target state are estimated).

Sex	minutes	speakers	sentences	phones
Male	1.09	6	16	923
Female	0.20	2	3	172
Total	1.29	8	19	1095

Table 3.7: The NCCCz evaluation subset statistics

3.2.4 SpeechDat-E Corpora

SpeechDat-E corpus consists of 5 East European languages which were collected via fixed telephone network. The corpus contains speech recording of Russian, Czech, Polish, Hungarian and Slovak languages and it is available via ELRA [106]. The corpus is well balanced with regards to age, gender and dialects and the number of speakers is between 1000-5000. Signals were recorded via fixed ISDN telephone network and sampled at 8 kHz and quantized using 8bit a-law format. These data were used for experiments for AF estimation and phone recognition tasks. Statistics of used corpora are summarized in Table 3.8.

corpora	set	speakers	gender (M/F)	sentences	hours
SPEECHDAT_CS	train	852	426/ 426	43139	73.4
	test	100	45/55	943	2.1
	dev	100	55/45	951	2.1
SPEECHDAT_SK	train	800	394/406	36526	49.56
	test	100	45/55	1117	1.88
	dev	100	55/45	1113	1.87
SPEECHDAT_HU	train	900	457/443	37861	51.2
	test	100	54/46	1085	1.73
	dev	100	54/46	1085	1.71
SPEECHDAT_PL	train	900	439/461	42506	69.64
	test	100	49/51	1148	2.3
	dev	100	52/48	1169	2.4
SPEECHDAT_RU	train	2000	993/1007	91079	126.2
	test	250	110/140	1827	3.55
	dev	250	139/111	1829	3.42

Table 3.8: SpeechDat-E data subsets

3.2.5 CtuTest - Czech read journal sentences

CtuTest is a private CTU database which consists of read journal sentences of various topics. The corpus contains 577 utterances from 40 speakers with the total duration of approximately 1 hour. Signals were recorded at 16 kHz sampling frequency and saved within 16bit linear-PCM format. This corpus was created mainly for purposes of AM evaluation.

3.2.6 CzLecDSP - Czech technical lectures

CzLecDSP is also a private CTU database which was collected within doctoral seminars containing technical lectures from the DSP field held at CTU [118]. The recorded data has spontaneous nature but they are more formal in comparison to NCCCz corpus. Signals were recorded at 16 kHz sampling frequency and saved within 16bit linear-PCM format. The corpora were mainly created for the purposes of evaluating continuous and more spontaneous speech.

Testing subsets			
database	speakers	utterances	hours
CtuTest	40	577	1.1
CzLecDSP	8	1417	1.7

Table 3.9: CtuTest & CzLecDSP data sets

3.2.7 TIMIT

The TIMIT database is commonly used for the evaluation of AF classification and phone recognition tasks for English by many authors [128], [32] as well as for other experiment related to speech recognition in general. It is used to compare our results of AF estimation with results obtained by other authors.

The speech data in TIMIT are recorded using a 16 kHz sampling frequency and manually labelled at phone level. Existing manual labels represent a significant benefit of this corpus and they were used in our experiments with the English language. Finally, the speech data without all SA utterances were chosen and divided into a standard subset such as training, cross-validation, and test one. A simplified set of 39 phones was used for experiments following the work of [128]. The contents of each subset are summarized in more details in Table 3.10.

data set	speakers	sentences	hours	num. words	num. boundaries
TRAIN	462	3696	3.14	30132	-
CORE test set	24	192	0.16	1570	7215
COMPLETE test set	168	1344	0.81	11025	50754

Table 3.10: TIMIT data sets used in presented evaluations

3.3 Evaluation criteria

The accuracy of LVCSR systems was measured on the basis of Word Error Rate (*WER*)

$$WER = \frac{S + D + I}{N} \times 100 \tag{3.1}$$

where N is total number of word from test list; S, D, and I are numbers of substitutions, deletions and insertions respectively, and Sentence Error Rate (*SER*) criteria

$$SER = \frac{C}{N} \times 100 \tag{3.2}$$

where N number of sentences in test set and C is the number of correctly decoded sentence.

When a performance of phone recognition systems was tested, Phone Error Rate (*PER*) was used for the evaluation. It was computed similarly to *WER* as

$$PER = \frac{S + D + I}{N} \times 100 \tag{3.3}$$

however, the numbers N, S, D, and I are related to number of all tokens, substitutions, deletions, and insertions at a phone level.

The quality of LMs used in our LVCSR systems was evaluated on the basis of two criteria *OOV* (Out-Of-Vocabulary words) and *PPL* (Perplexity). OOV quantifying the number of words which are not covered by a given LM (vocabulary) is computed as

$$OOV = \frac{U}{N} \times 100[\%], \tag{3.4}$$

where U is the number of unknown words and N is the total number of words in train set. $\widehat{\mathrm{P}}$PL, which describes a quality of LM, is defined as inverse normalized probability of word sequence $W = w_1 w_2 w_3 \ldots w_N$ given by test corpus and it can be approximated using a trained n-gram LM. The following equations is a specif case of the previous equation valid for a trigram LM

$$PPL(W) = 2^{LP(W)} \quad \left(\approx \frac{1}{\sqrt[N]{P(w_1 w_2 w_3 \ldots w_N)}} \right), \tag{3.5}$$

$$LP(W) = -\frac{1}{N} \sum_{i=1}^{N} \log_2 P(w_i | w_{i-2} w_{i-1}) \, . \tag{3.6}$$

This estimation was used in all experiments as well as for unigram and bigram LMs.

3.4 Particular results of the Czech ASR (LVCSR)

This section describes initial ASR experiments. The performance of ASR system under various acoustic and speaking style conditions is presented, i.e. ASR setups and achieved results for reading, spontaneous, and informal speech recognition are described. Particular parts of the mentioned ASR systems were used within experiments in this thesis and the presented results give an idea about the overall quality of these basic ASR modules as well as the whole system.

3.4.1 Setup of common ASR modules

The exact setup of key components of the basic Czech ASR system is described in this part, i.e. frond-end processing as well as acoustic and language modelling.

Frond-end processing

MFCCs features described in sections 2.1.1 and 2.2.1 were computed using CtuCopy tool with the following setup: preemphasis with the coefficient of 0.97 was applied, short-time frame had the length of 25 ms and it was moved with the step of 10 ms. Mel-filter bank contained 30 bands in the frequency range of 100-7940 Hz and 12 cepstral coefficients with additional $c[0]$ were computed. Cepstral mean normalization (CMN) was applied on per the speaker basis and these features were extended with delta and delta-delta parameters. LDA-based features for GMM-HMM architecture were based on static and normalized MFCC features extended with the a context of ±5 frames. The dimension of final feature vector was set to 40, see Fig. 2.4.

Acoustic modelling

Acoustic models were built using standard approaches for current ASR systems. The set of 45 Czech phones were expanded to the context-dependent crossword triphones. The set was extended with a silence phone which represented long silence. Concerning *GMM-HMM approach*, the initial context-independent AM (*mono*) consisted of left-right HMMs with 3 emitting states without skips for real non-silence phones and of 5 emitting states containing skip connections for silence phones. The *mono* alignment of the train set was used for building the context-dependent triphone-based AM (*tri1*). Phonetic decision tree and context-dependent phones were automatically derived using the data-driven approache. The *mono* and *tri1* AMs were trained using 13 MFCC coefficients and their delta, delta-delta coefficients. The next training process continued with the training of the second context-dependent AM (*tri2*) using the above mentioned LDA+MLLT features. The following step included training the *tri3* AM which used speaker adaptive training (SAT) using feature-space maximum likelihood linear regression (fMLLR). The derived fMLLR features and their delta, delta-delta were used for training the Subspace GMM (SGMM) system and the system was finally retrained discriminatively using bMMI criteria.

The topology of the *DNN-HMM hybrid approach* consisted of an input layer with 440 units (the context of 5 frames with 40 dimensional LDA-fMLLR features normalized with MVN) was followed by 6 hidden layers with 2048 neurons per layer and the sigmoid activation functions. The process of building of DNN-HMM system started with the initialization of hidden layers by Restricted Boltzmann Machines (RBMs) and then the output layer was added. The process continued by using the frame cross-entropy error function and ended with sMBR sequence-discriminative training.

Particular acronyms represent the following systems:

- “mono” - monophone GMM-HMM with MFCC features $+ \Delta + \Delta - \Delta$ features,
- “tri1” - triphone GMM-HMM with MFCC features $+ \Delta + \Delta - \Delta$ features,
- “tri2” - triphone GMM-HMM with LDA+MLLT features,
- “tri3” - triphone GMM-HMM with LDA+MLLT followed by SAT,
- “SGMM” - subspace GMM,

- “bMMI” - discriminatively trained models,
- “DNN” - cross-entropy trained DNN-HMM system.
- “DNN_sMBR” - discriminatively trained DNN-HMM system.

Language modelling

Concerning language modelling, we worked with standard n-gram-based statistical language models (LMs). The suitability of five general LMs for various speech recognition tasks was analyzed. The LMs were collected from three different publicly available resources, i.e. from Czech National Corpus (CNC) [51], Google n-grams distributed by Linguistic Data Consortium (WEB1T) [113], and from the corpora ORAL 2006, ORAL 2008, and ORAL 2013 produced by the Institute of Czech National Corpus [50]. General LMs from CNC and WEB1T corpora containing a general text were built in various numbers of word forms (60k, 120k, 340k) and the process of their creation is described in [113]. These models were expected to cover the general Czech language sufficiently.

3.4.2 Basic LVCSR Under Various Condition

This section presents the obtained results for the Czech ASR system under various acoustic and speaking styles conditions using two standard state-of-the-art architectures (GMM-HMM and DNN-HMM). The baseline recipes for the building of LVCSR using SpeechDat, SPEECON, CZKCC, and NCCCz corpora with the updated feature extraction tool CtuCopy which supports currently Kaldi format were analyzed. Obtained results are presented for whole AM training-cycle which started from *mono* AM, continued through *tri1*, *tri2*, *tri3*, *tri3_sgmm*, *tri3_sgmm_bmmi* systems and ended with *dnn*, *dnn_smbr* stages. The generic trigram CNC340k LM designed for the LVCSR task was used to present more realistic results.

Baseline LVCSR results for particular databases with matched LM

The performance of ASR systems, which were trained separately on particular databases is summarized in Table 3.11. It means that both acoustic and language models were trained using only the a train set of a particular database. The language models for these systems were built separately from a corpora of transcriptions contained in the train subsets. Bigram LMs were trained using Witten-Bell smoothing technique with the help of the SRILM toolkit [132]. We can observe the results between 10-44% WER depending on the system setup. The best result was achieved for SpeechDat setup and we can observe an increase of WER for SPEECON, more noisy car speech from CZKCC, and a serious increase of WER for spontaneous speech from NCCCz. Concerning DNN-HMM architecture, WER was reduced for all analyzed acoustic conditions in general, while the largest improvement was observed for sMBR discriminative technique.

		GMM-HMM						DNN-HMM	
	data set	mono	tri1	tri2	tri3	sgmm	bmmi	dnn	sMBR
SPEECON	test	24.36	17.52	16.90	16.86	15.78	15.43	15.00	13.73
office	dev	26.68	19.18	17.90	17.31	16.39	16.30	15.96	14.95
CZKCC	test	39.58	32.10	31.89	31.64	30.02	29.87	28.01	27.13
headset	dev	29.80	24.08	24.18	24.87	23.41	23.38	22.11	21.22
NCCCz	test	76.86	58.73	57.78	51.71	48.40	46.84	46.65	43.45
SPEECHDAT	test	22.46	14.33	14.36	14.78	14.05	13.97	13.04	10.86
Czech	dev	20.06	13.96	13.96	14.08	13.44	13.32	13.28	11.15

Table 3.11: Baseline results for particular databases with matched LM

Results for LVCSR using CNC language model

The above described results were achieved using an optimal setup and thus report on an ideal case. The main issue was that the LMs were created from available transcriptions that could potentially contain texts from the testing sets of SpeechDat, SPEECON, and CZKCC databasest since the prompt sheets used for recording were not completely disjunct among speakers.

Table 3.12 summarizes the results for the LVCSR task using 340k-word language model created from the Czech National Corpus (CNC340k LM). It is possible to observe an increase in WERs in comparison to the results in Table 3.11. Slightly higher WERs were most likely due to the fact that test set utterances contained phonetically rich sentences with a slightly enhanced appearance of words with rare phones. The contribution of more complex AMs is clearly apparent when we compare the results of context-independent *mono* vs. context-dependent *tri1*, speaker-independent *tri2* vs. speaker-dependent *tri3* acoustic models, or discriminatively trained GMM-HMM *bmmi* vs. discriminatively trained DNN-HMM *dnn_smbr*. The AMs based on *mono*, *tri1*, *dnn* were trained on SPEECON corpus and were later analyzed using experiments focusing on phonetic segmentation. The speaker dependent AM *tri3* were used for generating frame alignment within experiments with AF and phone recognition.

Results for LVCSR under far field acoustic conditions

The following experiments focused on evaluating the ASR system under more adverse acoustic conditions. Channels other than SPEECON office were used for this evaluation.

		GMM-HMM						DNN-HMM	
	data set	mono	tri1	tri2	tri3	sgmm	bmmi	dnn	sMBR
SPEECON office	test	51.32	31.37	27.48	23.54	19.59	19.06	18.99	17.46
CZKCC headset	test	29.20	15.90	14.79	11.57	9.56	9.43	9.39	8.4
NCCCz	test	88.09	69.32	66.29	59.92	57.63	55.63	51.15	48.79

Table 3.12: The results for particular databases using CNC340k LM

		GMM-HMM					
	data set	mono	tri1	tri2	tri3	sgmm	bmmi
SPEECON CS0	test	53.85	29.83	27.49	22.08	19.42	18.34
	dev	51.82	28.12	25.36	20.40	17.33	16.49
SPEECON CS1	test	63.63	36.92	35.82	27.95	23.90	23.21
	dev	63.07	32.76	24.46	22.00	20.74	20.36
SPEECON CS2	test	67.59	39.59	37.73	30.33	26.77	24.39
	dev	72.13	44.29	40.86	30.40	27.05	24.24
SPEECON CS3	test	92.85	77.72	74.56	67.57	61.23	58.59
	dev	95.38	84.39	81.06	74.80	68.60	65.46

Table 3.13: The results for all channels in SPEECON using CNC340k LM

The impact of far field microphones CS1, CS2 and CS3 channels was apparent by looking at the obtained results. The absolute difference between CS0 and CS1 and CS2 is around 6% WER for SNR within the of range of $19 - 13dB$. The performance on channel CS3 was above 50% WER and the corresponding SNR was around $6dB$. The channel CS3 was also found to be challenging in the phonetic alignment task, the most likely causes being the far field speech nature and a very low volume of recorded audios.

Partial conclusions

The above described results prove that the developed LVCSR system for Czech can achieve stat-of-the-art accuracy comparable to other ASR systems. This observation held true for various acoustic conditions and speaking styles and proved that the trained AMs could be used for further research which focused mainly on AF and phonetic segmentation. Some results are also used as baseline results in the following experiments.

3.4.3 ASR framework for Causal Czech Recognition

The purpose of this section is to extend on the previous baseline results for NCCCz and to describe the design of more sophisticated ASR system for Czech casual speech recognition task. The focus was on the contributions of acoustic and language models as well as on pronunciation lexicon optimization. The AM was trained on large speech train set which consists of several Czech corpora available at our department. Special attention was also paid to the impact of publicly available corpora suitable for LM creation.

The section starts with discussion about the state-of-the-art of Czech Casual Speech recognition and continues with the description of implemented solutions to improve the accuracy of casual speech recognition. It is divided into three subsections: robust acoustic modelling, improvement to language modelling and extensions of pronunciation lexicons. Results of particular experiments are discussed in the context of other results obtained for other speaking styles. The section also presents a comparison between the GMM-HMM system and DNN-HMM hybrid approach.

The recognition of spontaneous speech still represents a very challenging task. The commonly achieved accuracy is still rather low in comparison with a generally high accuracy for standard LVCSR systems. This conclusion is supported by many other works for other languages [70, 6, 95, 97, 116, 20, 129]. The spontaneous or colloquial speech recognition deals with similar problems, i.e. strong variability in the pronunciation (mainly strong pronunciation reduction), changes in the word morphology, free word order in the sentence, sentence breaks, and some others [75, 94].

Many authors have presented solutions for the above mentioned problems and achieved varying results for various languages, speaking styles, or recording conditions. The authors in [13] worked with transcriptions of oral interviews of survivors and witnesses of the Holocaust and they reported 39.60% WER for English and 39.40% for Czech. However, when the level of speech spontaneity is higher, typically for very informal speaking style, the accuracy of speech recognition falls. Authors in [70] worked with the recordings of telephone conversations and reported 48% WER for the Czech language. Similarly in [97], authors presented results around 31-56% WER for the case of a very informal speech recognition task. Results presented by other authors were also confirmed also by our evaluation of casual speech recognition which were based on data from NCCCz, described previously. The recognition accuracy in a standard LVCSR task using a standard setup decreased significantly, see Table 3.11 and Table 3.12. The possible improvement of these results is discussed in the following parts this section.

Impact of front-end processing & Acoustic modelling

The front-end processing and AM training for NCCCz followed the setup previously described in 3.4.1. This was possible mainly because the conversations available in NCCCz were recorded in a quiet environment which was similar to headset recordings from a quiet SPEECON environment. Other speech corpora which were similar to NCCCz, from an acoustic conditions point of view, were also included in order to create a larger and more generic train set. This was especially important for DNN-HMM system. To summarize, office subpart of SPEECON (SPEECON_CS0_OFFICE), clean subpart of car database (CZKCCC_headset) and training part of NCCCz (NCCCz_train) were used as a set for AM training in all further experiments.

Impact of language models for casual Czech

The standard n-gram-based statistical LMs described in section 3.4.1 were used for NCCCz corpus. With regards to NCCCz corpus, the significant problem which had to be solved was a choice of a suitable resources that would appropriately cover the casual speech. The suitability of five general LMs collected from three different publicly available resources, CNC, WEB1T, ORAL_2006, ORAL_2008 and ORAL_2013 were analyzed. While the corpora CNC, same as WEB1T, contained text that was rather general in nature that were built with various size of word forms up-to 340k and these models should cover general nature of Czech. The corpora of ORAL family contain spontaneous conversations and it was thus expected the produced LM would be a better fit for the NCCCz domain. The number of word forms obtained for ORAL corpus was 162k and 29k for NCCCz. This differences amounted to 73k additional words from ORAL and 9k words from NCCCz approximately. Finally, in order to cover the maximum vocabulary for our task, we have also created LMs from NCCCz. The first LM was trained from a defined training part of NCCCz containing the transcription of 60% utterances per each recorded session which were also not used for the evaluations later. It represented a slightly more realistic scenario as the content of recognized utterances has not been seen before. The second LM was created for comparison purposes as an optimal LM for causal speech since it was made from all available NCCCz transcriptions.

Impact of pronunciation variation modelling

The modelling of pronunciation variation in casual speech (mainly pronunciation reductions) was the last point of interest. The particular rules, some of them known from other works, e.g. [94] or [127], others obtained from results of the psycholinguistic study of pronunciation reduction in NCCCz [69] were applied. In the end, we have used approximately

6700 additional pronunciation variants. The illustrative examples of several rules are

“v[sSzZ]→[sSzZ]” - e.g. “*vždyt’, vstát’*” (“*but, to stand up*”),

“[td]J → [cJ\J]” - e.g. “*letní*” (“adj. *summer*”),

“cons_1-t-cons_2 → cons_1-cons_2” - e.g. “*jestli*” (“*if*”),

“js → s” - e.g. “*jsem*” (“*I am*”),

“j[eai] → [eai]” - e.g. “*jestli, jinam*” (“*if, elsewhere*”),

“zj → z” - e.g. “*zjistíš*” (“*You will find*”),

“t-S → t_S” - e.g. “*většina*” (“*majority*”),

“nsk → nt_sk” - e.g. “*č]ínský*” (“*Chinesse*”),

“vZd → vd” - e.g. “*vždycky*” (“*always*”).

Results of experiments & discussion

The achieved results for previously established recognition tasks are evaluated from the following points of view: the *optimization of acoustic modelling*, the impact of *language modelling* and *pronunciation variation*. Experiments were performed on utterances from the following Czech databases: SPEECON, CtuTest, CzLecDSP, and NCCCz which cover different levels of spontaneity, i.e.

- *T1 - read speech recognition*
  a) read sentences, phonetically rich (SPEECON database),
  b) read journal sentences, phonetically unbalanced (CtuTest database),

- *T2 - spontaneous speech recognition*
  recordings of technical lectures (CzLecDSP database),

- *T3 - casual speech recognition*
  recordings of highly informal conversations (NCCCz database).

The principal results of these experiments are those for spontaneous speech data from NCCCz and CzLecDSP (test sets TA2 and TA3). Experiments performed on testing subsets from SPEECON and CtuTest (test sets T1a and T1b) which contained read speech were done for comparison purposes to analyze the overall recognizer setup in a more standard task.

I. The impact of AM type

The first results describe the quality of used AM, i.e. starting from a basic GMM-HMM approach and ending with the best AM based on a DNN-HMM architecture. General 340k-word bigram LM based on CNC was used for all of these experiments. The obtained results shown in Table 3.14 demonstrate that our DNN-HMM LVCSR system obtained accuracy comparable to the current state-of-the art systems, i.e. 15.2% of WER for standard read speech. For spontaneous speech we have obtained WER of 37.4% for the task of lecture transcription (i.e. with slightly more formal speaking style) and 72.0% for very informal (casual) speech from NCCCz.

tasks	tri2	tri3	SGMM	bMMI	DNN
T1a	29.8	23.4	22.2	21.8	21.1
T1b	24.0	17.0	15.9	15.3	15.2
T2	49.9	41.3	39.9	38.0	37.4
T3	82.5	76.1	74.9	74.2	72.0

Table 3.14: WERs of LVCSR in the phase of AM optimization

II. The impact of LM

Results shown in Table 3.15 present the analysis of various LMs. The first part summarizes achieved WERs for all speaking styles using general CNC and WEB1T-based LMs where the strong decrease for the case of casual speech is clearly shown. The second part of Table 3.15 presents the results for TA3 task (casual speech) and using LMs trained on ORAL and NCCCz (i.e. transcriptions of recorded casual speech). The reduction of out-of-vocabulary (*OOV*) rate as well as the perplexity (*PPL*) confirmed improved match for casual speech and resulted in *WER* of around 60-70%. The achieved results also showed that trigram-based LMs brought a very small improvement in *WER* but the complexity of used HCLG graph increased significantly. Due to this fact, bi-gram LMs were used in further experiments. The last line of Table 3.15 represented a rather exceptional case where the LM *NCCCzAll* was created from all available transcriptions in NCCCz (i.e. including also the test set). This model had *OOV* of 0% and a very low value of *PPL*, both of which were expected. This result was presented purely as a limit case to demonstrate the theoretical limits of used modelling approaches.

The next experiments were focused on minimizing OOV and WER in the TA3 task by merging of various bigram LMs. The results for merged LMs with the uniform interpolation weights are summarized in Table 3.16. The usage of various merged LMs reduced the level of OOV significantly but the WER decreased only marginally as the setup of the interpolation weights (λ) was not optimal. Therefore, we also optimized the value of

Tasks	*LM*	OOV	PPL	*2-gram*	3-gram
TA1a	CNC	1.6	3572	21.1	21.8
TA1b	CNC	1.8	2034	15.2	14.7
TA2	CNC	4.8	2937	37.4	37.2
TA3	CNC	4.6	2065	72.0	72.2
	WEB1T	4.5	4427	68.9	-
TA3	ORAL06	6.5	389	67.1	66.4
	ORAL08	6.7	445	66.8	66.3
	ORAL13	4.7	475	66.1	65.4
	ORALall	4.0	426	63.6	62.5
	NCCCz60	7.2	248	61.4	61.2
	NCCCzAll	*0*	*69*	*41.3*	*28.4*

Table 3.15: WERs of LVCSR with various 2-gram a 3-gram LMs on particular tasks.

bigram LMs	*OOV*	*WER*
CNC+WEB1T	4.3	69.8
CNC+WEB1T+ORALall	2.8	64.7
CNC+WEB1T+ORALall+NCCCz60	1.5	61.2

Table 3.16: DNN-HMM casual speech recognition (TA3) with merged bigram LMs.

		NCCCz weight λ				
LMs	OOV	0.0	0.25	0.50	0.75	1
CNK340+NCCCz60	2.2	72.0	62.8	60.8	59.4	61.4
ORALall+NCCCz60	2.5	63.6	60.9	59.8	58.9	61.4
WEB1T+NCCCz60	2.1	68.9	62.3	60.6	60.0	61.4

Table 3.17: DNN-HMM with various weights of NCCCz in merged LMs on TA3 task.

λ for particular LMs. The best result was obtained with the following weights λ: 0.2 for ORAL LM, 0.15 for CNC 0.15 for WEB1T and 0.5 for NCCCz. The corresponding WER reached about 59.7%. The final investigation focused on merging various LMs with the NCCCz-based LM. The contributions of various interpolation weights λ to the final WER are summarized in Table 3.17. The best results were achieved for the setup with $\lambda = 0.75$.

In the end, the combination of all LMs brought an improvement in target OOV but the decrease of WER was smaller. The results proved that general LMs (CNC and WEB1T) did not contain proper information to describe the causal speech in NCCCz. However, the LMs created from ORAL corpus modelled casual speech very similarly to a LM created directly from NCCCz, with the exception of NCCCzAll language model used also the test data.

III. The impact of pronunciation reduction

The final results presented in this chapter describe the achieved WER for three approaches of pronunciation modelling. First, automatically generated pronunciation was used for all words in analyzed LMs (which is used always if a word is not present in the available dictionary). Second, an approved canonic pronunciation of all words from NCCCz was created by manually by two independent experts. Third, the dictionary with the additional pronunciation variants containing phone reductions using the above-described rules was used. All obtained results are summarized in Table 3.18 and, according to preliminary assumptions, the recognition accuracy has improved but by only about 1.4%.

LM	*Lexicon*	*WER*
0.25 ORALall + 0.75 NCCCz60	automatic	59.8
	canonic checked	58.9
	reduction variants	58.4

Table 3.18: Impact of pronunciation variation in DNN-HMM system

Conclusions

This section describes an optimization of DNN-HMM and GMM-HMM based LVCSR for casual speech recognition for Czech and its performance on data from the Nijmegen Corpus of Casual Speech. Achieved results confirmed that it is possible to use these systems for casual speech recognition, but the results are significantly worse when compared to the results for more formal speech. It was also proved that publicly available corpora ORAL which contains transcriptions of spontaneous conversations and corpora of formal Czech can be used for the creation of basic LMs for the task of casual speech recognition.

Chapter 4

Estimation of AF for Czech and other languages

This chapter summarizes the research on the estimation of AF from an acoustic speech signal. The term of AF is introduced and the definition of AF classes for Czech, English, and several other languages is discussed. Further, widely used approaches of AF estimation are summarized. The chapter is closed by a description of performed analyzes of AF estimation realized for particular languages as well as acoustic conditions.

4.1 Articulatory features for analyzed languages

The term it Articulatory features generally represents a set of features trying to describe how the human speech is generated. Articulatory information can be obtained using direct measurements of the motion of particular articulators (e.g. lips, tongue, jaw) or various statistic methods estimating this information from the acoustic speech signal. Since a lot of approaches to achieve articulatory information it have been suggested, there are various ways to represent AF.

With regards to the statistical methods, the representations of AF are standardly based on articulatory phonetics or different theories of phonology [60], [78]. The following three representations of AF are the most important ones:

- multi-valued features which are based on articulatory phonetic categories,
- phonological distinctive features proposed by Chomsky and Halle,
- articulatory gestures used in articulatory phonology and proposed by Browman and Goldstein.

AF based on multi-valued features or articulatory gestures are widely applied in the speech applications which were previously mentioned. They are commonly used for observation

AF class	*Cardinality*	*Feature values*
English		
place	10	alveolar, dental, labial, postalveolar, rhotic, velar, labiodental, lateral, none
degree	6	approximant, closure, flap, fricative, vowel
nasality	3	front, central, back
rounding	3	stop, nasals, affricates, fricatives
glottal sta.	4	aspirated, voiceless, voiced
vowel	23	aa, ae, ah, ao, aw1, aw2, ax, ay1, ay2, eh, er, ey1, ey2, ih, iy, ow1, ow2, oy1, oy2, uh, uw, nil
height	8	high, low, mid, mid-high, midlow, very-high, nil
frontness	7	back, front, mid, mid-back, mid-front, nil
voicing	3	voiced, unvoiced

Table 4.1: AF classes for English

modelling in ASR, robust speech recognition, and nowadays also in important areas of multilingual/cross-lingual ASR or low-resource speech recognition. In the contrast, the AF based on articulatory gestures are standardly used in the task of pronunciation modelling. These different representations of AF were analyzed separately in several experiments during Johns Hopkins University (JHU) Summer Workshop [78] and it was also proposed how these AF sets could be combined in ASR system by Hasegawa-Johnson [45].

This work analyzes the AF-based TANDEM approach, which was presented in JHW [33], [14], [72] with focus on Czech language. Therefore, it uses AF based on multi-valued features (further referred to only as AF) for observation modeling with the aim to improve general ASR accuracy, phone recognition, as well as phonetic segmentation precision for the analysis of Czech spontaneous speech.

AF with multi-valued feature representation of speech production knowledge for observation modelling was used with the purpose of making these features acoustically distinguishable which is discussed more within next sections.

4.1.1 AF set for English

As mentioned previously, AF are principally defined on the basis of a particular phone generation (articulation) which deals with the articulatory phonetics. When multi-valued features are defined it is commonly proceeded on the basis of the International Phonetic Alphabet which divides distinctive sound to phonetics categories such as manner, place, voicing and others. For English, several approaches of multi-valued features definition are used with slightly varying amount of classes and categories; i.e. defined on JHU Summer Workshop [78] for AFs classification (JHU set), for better uses in the case of manual transcriptions [79], or in the task of automatic phonetic segmentation (Hosom set) in [49]. Hosom marks these features as distinctive phonetic features. In [120], authors compared two approaches based on JHU or Hosom sets. Their results showed that both approaches could achieve similar classification accuracy.

Finally, within this work the JHU approach with one additional class for the voicing [33], [17] is used. A brief overview of used AF for English is presented in Table 4.1. More details on phone mapping to AF can be found in [33].

4.1.2 AF set for Czech

For the Czech language, AF have not yet been defined unambiguously. Therefore, it was necessary to define them similarly to the above-mentioned English standard, taking into account the phone categories used standardly for Czech [142]. Standard inventory of phones for Czech defined by SAMPA standard [143] consists of 49 phones including several rare allophones as well as schwa and glottal stop which do not appear in Czech canonical pronunciation.

Within this work, the same set of phones which was standardized for Czech ASR systems was used. This set does not contain syllabic variants of consonants, i.e. phones "*m=*, *l=*, *r=*", and voiced phone "*G*" which appears only in very special contexts at word boundaries, as well as glottal stop "*?*" which also does not have regular appearance in Czech pronunciation.

The resulting phone inventory consists of 44 phones (including diphthongs) which can be categorized into the following phonetic classes according to the methodology described for English in [66], [17], [61] together with the application of standard conventions for Czech defined by [100] and [142]. The more particular details of Czech vowel and consonant categorization are described Table 4.2 and 4.3 and final multi-valued features-based AF for the Czech language are then summarized in Table 4.4. Each AF class is completed by the 'silence' value which increases class cardinality. The phones which cannot be put into categories within a particular AF class (e.g. vowel *'a'* is not eligible for consonants categories) are marked by the value 'nil'. The complete overview of the AF used for the complete Czech phone inventory is shown in Table 4.5.

4.1.3 AF sets for Speechdat-E languages

To prove the language independence of an estimation of AF, the research on other languages was also performed. Finally, Slovak, Polish, Hungarian, and Russian have been selected as languages of SpeechDat-E corpora set which is available at our department. The AF for these languages have not been defined unambiguously to our best knowledge. Therefore, this section provides a summary of AF definition for given languages.

Similarly to Czech, the multi-valued AF based on phone inventory mapping to partic-

			Manner											
			stop			*affricates*		*fricatives*				*approximants*		
			plosives		*nasals*					*trills*			*lateral*	*glides*
Place	*labial*	*bilabial*	*p*	b	m									
		labiodental			ɱ			f	v					
	alveolar	*prealveolar*	*t*	d	n	ts	dz	s	z	r̝̊	r̝	r	l	
		postalveolar				*tʃ*	tʒ	ʃ	ʒ					
	palatal		*c*	ɟ	ɲ									j
	velar		*k*	g	ŋ			x						
	glottal								ɦ					
Sonority (Sonors/Noises)			No	No	So	No	No	No	No	No	No	So	So	So
Voicing (Voiced/Unvoiced)			U	V	V	U	V	N	V	U	V	V	V	V

Table 4.2: Phonetic categorization of Czech consonants

		Manner		
		front	*central*	*back*
Place	*close*			u
	close-mid	ɪ	ə	o
	open-mid	ɛ		
	open		a	
Rounding		*unrounded*		*rounded*

Table 4.3: Phonetic categorization of Czech vowels

AF class	*Cardinality*	*Feature values*
Voicing	3	voiced, unvoiced
Place_con	9	bilabial, labiodental, prealveolar, postalveolar, palatal, velar, glottal, nil
Place_vow	5	front, central, back, nil
Manner_con	9	stop, nasals, affricates, fricatives, trills, lateral, glides, nil
Manner_vow	5	open, mid, close, nil
Rounding	4	rounded, unrounded, nil
Sonority	4	noise, sonor, nil

Table 4.4: AF classes for the Czech language

ular articulatory-phonetic classes using the IPA table were defined. To assign phones to their articulatory categories, the phonetic inventory unification was to be first involved to define mapping from SAMPA (Speech Assessment Methods Phonetic Alphabet) to IPA because SAMPA alphabets which were used to represent phonetic transcription in dictionary did not use systematically same symbols for all equivalent phones in particular languages. This phonetic inventory unification was defined in [28] and within this thesis, the unification of AF classes and mapping phones to articular categories for all languages

Phones	*Voicing*	*Place_con*	*Place_vow*	*Manner_con*	*Manner_vow*	*Round*	*Sonor*
ɪ	+	nil	front	nil	high	−	nil
ɛ	+	nil	front	nil	middle	−	nil
a	+	nil	central	nil	low	−	nil
o	+	nil	back	nil	middle	+	nil
u	+	nil	back	nil	high	+	nil
ɪ*:*	+	nil	front	nil	high	−	nil
ɛ*:*	+	nil	front	nil	middle	−	nil
a:	+	nil	central	nil	low	−	nil
o:	+	nil	back	nil	middle	+	nil
u:	+	nil	back	nil	high	+	nil
o_u	+	nil	back	nil	middle	+	nil
a_u	+	nil	central	nil	low	−	nil
ɛ*_u*	+	nil	front	nil	middle	−	nil
ə	+	nil	central	nil	middle	nil	nil
p	−	bilabial	nil	stop	nil	nil	−
b	+	bilabial	nil	stop	nil	nil	−
t	−	prealveolar	nil	stop	nil	nil	−
d	+	prealveolar	nil	stop	nil	nil	−
c	−	palatal	nil	stop	nil	nil	−
ɟ	+	palatal	nil	stop	nil	nil	−
k	−	velar	nil	stop	nil	nil	−
g	+	velar	nil	stop	nil	nil	−
tʃ	−	prealveolar	nil	affricates	nil	nil	−
dʒ	+	prealveolar	nil	affricates	nil	nil	−
tʃ	−	postalveolar	nil	affricates	nil	nil	−
dʒ	+	postalveolar	nil	affricates	nil	nil	−
f	−	labiodental	nil	fricatives	nil	nil	−
v	+	labiodental	nil	fricatives	nil	nil	−
s	−	prealveolar	nil	fricatives	nil	nil	−
z	+	prealveolar	nil	fricatives	nil	nil	−
r̝	−	prealveolar	nil	trills	nil	nil	−
r̝̊	+	prealveolar	nil	trills	nil	nil	+
ʃ	−	postalveolar	nil	fricatives	nil	nil	−
ʒ	+	postalveolar	nil	fricatives	nil	nil	−
j	+	palatal	nil	glides	nil	nil	+
x	−	velar	nil	fricatives	nil	nil	−
ɦ	−	glottal	nil	fricatives	nil	nil	−
r	+	prealveolar	nil	trills	nil	nil	+
l	+	prealveolar	nil	lateral	nil	nil	+
m	+	bilabial	nil	nasals	nil	nil	+
n	+	prealveolar	nil	nasals	nil	nil	+
ŋ	+	velar	nil	nasals	nil	nil	+
ɲ	+	palatal	nil	nasals	nil	nil	+
ɱ	+	labiodental	nil	nasals	nil	nil	+

Table 4.5: Summary of articulatory features per particular Czech phones

is completed.

Concerning particular languages, a 50 phones set with 10 vowels, 4 diphthongs and 36 consonants was defined finally for Slovak. The Slovak phones set consists of the long lateral *l:* and long trill *r:* which represent allophones of *l/r* phones. As it is mentioned in the [44], these allophones appear in very special contexts and with regards to the frequency of these allophones in Slovak SpeechDat corpus, 0.07% for *r:* and 0.09% for *l:*, we decided to map the allophones to the same AF class as *l/r* phones. The categorization of the Slovak phones according to the position in IPA table is summarized in the Table 4.6 for consonants and in the Table 4.7 for the vowels.

The Polish language consists of 37 phones with 28 consonants and 9 vowels. The categorization of the Polish phones was defined with regards to the Polish IPA reference in [53] and it is summarized in the Table 4.8 for consonants and in the Table 4.9 for the vowels. Hungarian, as the phonetically richest language, consists of 68 phones with 54 consonants and 14 vowels. The conversation was realized based on description of the Hungarian IPA in the [133]. The categorization of the Hungarian phones is summarized in the Table 4.10 for consonants and in the Table 4.11 for the vowels. Finally, the Russian language consists of 50 phones with 38 consonants and 12 vowels. To distinguish the Russian consonants, the palatalization class has to be involved. The categorization of the Russian phones was defined based on Russian IPA description in the [147] and is summarized in the Table 4.12 for consonants and in the Table 4.13 for the vowels. The complete overview of articulatory features for complete phone sets per particular Slovak, Polish, Hungarian, Russian is available in Appendix A.

			Manner											
			stop			*affricates*		*fricatives*				*approximants*		
			plosives		*nasals*					*trills*			*lateral*	*glides*
Place	*labial*	*bilabial*	*p*	b	m									
		labiodental			ɱ			f	v					
	alveolar	*prealveolar*	*t*	d	n	ts	dz	s	z			r	l	
		postalveolar				tʃ	tʒ	ʃ	ʒ					
	alveolopalatal					tɕ	dʑ	ɕ	ʑ					
	palatal		*c*		ɲ								ʎ	j
	velar		*k*	g	ŋ			x						
	glottal								*h*					
Sonority (Sonors/Noises)			No	No	So	No	No	No	No	No	No	So	So	So
Voicing (Voiced/Unvoiced)			U	V	V	U	V	N	V	U	V	V	V	V

Table 4.6: Phonetic categorization of consonants for SK

		Manner		
		front	*central*	*back*
Place	*close*	i		u
	close-mid	e	ə	o
	open-mid			
	open		a	
Rounding		*unrounded*		*rounded*

Table 4.7: Phonetic categorization of vowels for SK

			Manner											
			stop			*affricates*		*fricatives*				*approximants*		
			plosives		*nasals*					*trills*			*lateral*	*glides*
Place	*labial*	*bilabial*	*p*	b	m									
		labiodental						f	v					
	alveolar	*prealveolar*	*t*	d	n	ts	dz	s	z			r	l	
		postalveolar				tʃ	tʒ	ʃ	ʒ					
	alveolopalatal					tɕ	dʑ	ɕ	ʑ					
	palatal				ɲ									j
	velar		*k*	g	ŋ			x						
	glottal													
Sonority (Sonors/Noises)			No	No	So	No	No	No	No	No	No	So	So	So
Voicing (Voiced/Unvoiced)			U	V	V	U	V	N	V	U	V	V	V	V

Table 4.8: Phonetic categorization of consonants for PL

		Manner		
		front	*central*	*back*
Place	*close*	i	ɨ	u
	close-mid	e	ə	o
	open-mid			
	open		a	
Rounding		*unrounded*		*rounded*

Table 4.9: Phonetic categorization of vowels for PL

			Manner											
			stop			*affricates*		*fricatives*				*approximants*		
			plosives		*nasals*					*trills*			*lateral*	*glides*
Place	*labial*	*bilabial*	p	b	m									
		labiodental			ɱ			f	v					
	alveolar	*prealveolar*	t	d	n	ts	dz	s	z			r	l	
		postalveolar				tʃ	tʒ	ʃ	ʒ					
	alveolopalatal					tɕ	dʑ	ɕ	ʑ					
	palatal		c	ɟ	ɲ			ç						j
	velar		k		ŋ			x						
	glottal								h,ɦ					
Sonority (Sonors/Noises)			No	No	So	No	No	No	No	No	No	So	So	So
Voicing (Voiced/Unvoiced)			U	V	V	U	V	N	V	U	V	V	V	V

Table 4.10: Phonetic categorization of consonants for HU

		Manner		
		front	*central*	*back*
Place	*close*	i, y		u
	close-mid	ø	ə	o
	open-mid	ɛ		
	open			ɑ
Rounding		*unrounded*		*rounded*

Table 4.11: Phonetic categorization of vowels for HU

			Manner											
			stop			*affricates*		*fricatives*				*approximants*		
			plosives		*nasals*					*trills*			*lateral*	*glides*
Place	*labial*	*bilabial*	p	b	m									
		labiodental			ɱ			f	v					ʋ
	alveolar	*prealveolar*	t	d		ts		s	z			r	l	
		postalveolar						ʃ	ʒ					
	alveolopalatal					tɕ	dʑ	ɕ	ʑ					
	palatal		c	ɟ				ç						j
	velar		k	g	ŋ			x						
	glottal													
Sonority (Sonors/Noises)			No	No	So	No	No	No	No	No	No	So	So	So
Voicing (Voiced/Unvoiced)			U	V	V	U	V	N	V	U	V	V	V	V
palatalization			p^j, b^j, f^j, v^j, r^j, l^j, $t\int^j$, $\int^j$, k^j, g^j											

Table 4.12: Phonetic categorization of consonants for RU

		Manner		
		front	*central*	*back*
Place	*close*	i	ɨ	u
	close-mid	e	ə	o
	open-mid			
	open		a	
Rounding		*unrounded*		*rounded*

Table 4.13: Phonetic categorization of vowels for RU

4.2 AF Estimation techniques

In this section, we summarize the widely used approaches for estimation of AF features from acoustic signal. The results of experiments focusing on the analysis of common acoustic features in the task of Czech AF classification are presented. It follows with experiments focusing on usage of temporal context for AF estimation. Two different approaches are analyzed. The experiments for the Czech language end with analyzes of the robust AF estimation under various environments or channels including different types of noise. The section 4.2.4 continues with estimation of AF for English languages. The suitability of other types of ASR features for the AF estimation is discussed and the optimization of DNN hyper-parameters is analyzed. Finally, the chapter ends with a review of AF estimation for other languages and for telephone acoustic conditions.

AF estimation techniques

Various machine learning algorithms have been used for the estimation of articulatory features. The Artificial Neural Network (ANN) [33], [120] and the Dynamic Bayesian Network are among the most frequently used approaches. Also other classifiers such as the Hidden Markov models (HMM), *k*-nearest neighbour algorithm, Gaussian Mixture Model (GMM) or classifiers using multi-task learning are used [61], [120], [93]. Deep Neural Networks are successfully used for AF estimation in works [48], [148], [99]. Nowadays, the end-to-end are becoming very popular in the speech community generally and they are also used by some authors for AF estimation with very promising results [59].

Therefore, in this work, both the MLP with one hidden layer and DNN structures with more hidden layers for particular AF classes were analyzed. The size of the output layer is always given by the cardinality of the estimated AF class, as well as the size of the input layer is given by the size of the input speech features. The size of the hidden layer is typically set experimentally for the particular AF class.

AF classification accuracy

The accuracy of AF classification is typically measured on the level of percentage of correctly recognized frames, i.e. Frame Accuracy (*FAcc*) defined as

$$FAcc = \frac{n\_correct\_frame\_labels}{total\_frames} \cdot 100\,. \qquad (4.1)$$

4.2.1 MLP-based AF estimation with common acoustic features

This section deals with the usage of basic acoustic features, i.e. Mel-frequency cepstral coefficients (MFCC) or Perceptual Linear Prediction (PLP) coefficients, in the task of MLP-based estimation of AFs for the Czech language. These features describe typically the short-time spectral representation of a speech signal at the input MLP network [68].

Other representations could be used as well, e.g. RASTA-PLP, spectrum derivative feature (SDF), linear predictive coding coefficients(LPC). More detailed comparative study of various acoustic features is reported in [35].

Realized analysis is focused on the study of the contribution of dynamic features such as delta, double-delta, as well as triple-delta coefficients to the resulting accuracy of AF classification. Generally, it is known that the expansion of static features joined with their temporal derivatives improves the overall accuracy of the task of speech recognition [71]. These features carry the most simple way the time context information of a signal, i.e. they give an information about trajectories of static features over time [62] so it was used also for the estimation of AF using basic MLP network.

Experimental setup

MFCC/PLP features were computed using the following exact setup:

- MFCC
  - preemphasis coefficient of 0.97,
  - Hamming window with length of 25 ms,
  - window shift of 10 ms,
  - 30 filters in auditory based spectral analysis,
  - 12 cepstral coefficients with the additional zeroth cepstral coefficient,
- PLP
  - Hamming window with length of 25 ms,
  - window shift of 10 ms,
  - 20 filters in PLP-based auditory filter bank (for 16 kHz speech data),
  - 12 cepstral coefficients with the additional zeroth cepstral coefficient,

The total length of these features vector varied from 13 to 52 coefficients. The subpart from the Czech SPEECON database marked as the OFFICE set was used for all experiments. The sizes of train, CV and test sets are summarized in Table 3.2 and they are described in more details in section 3.2.1. The results were measured on the basis of *FAcc*.

Results & Discussion

I. Impact of dynamic features

The first analysis was focused on the usage of various dynamic features (Δ, $\Delta\Delta$, $\Delta\Delta\Delta$). Detailed results of this experiments are shown in Fig. 4.1 and 4.2. All temporal derivative variants of static features significantly improved the MLP based AF classification as it can be seen from bar graphs in Fig. 4.1 and 4.2. To analyze the contribution of temporal

	MFCC				*PLP*			
	0	*0_d*	*0_d_a*	*0_d_a_t*	*0*	*0_d*	*0_d_a*	*0_d_a_t*
Voicing	*0*	3.6	5.2	6.1	*0*	3.6	5.3	6.2
Place_con	*0*	10.0	13.1	14.6	*0*	9.4	13.0	14.4
Place_vow	*0*	9.3	11.7	12.8	*0*	8.9	11.2	12.6
Manner_con	*0*	10.4	14.3	15.7	*0*	10.4	14.1	15.8
Manner_vow	*0*	8.5	11.2	12.2	*0*	8.6	11.3	12.5
Rounding	*0*	8.89	11.5	12.2	*0*	8.6	10.9	12.3
Sonor	*0*	9.1	12.3	13.4	*0*	9.2	12.3	13.4
avg	*0*	8.5	11.3	12.4	*0*	8.4	11.1	12.4

Table 4.14: The absolute improvement of AF estimation accuracy.

derivatives, the absolute improvement of *FAcc* during the expansion of static features by their temporal derivatives is also presented in Table 4.14.

The best results for both features used (MFCC and PLP) were obtained when all differential parameters (up to the 3rd derivative) were used. However, the setup of feature extraction with *0_d_a_t* coefficients achieved only a little absolute improvement (across all AF classes about 1.1% for MFCC and about 1.3% for PLP) compared to the setup with *0_d_a* coefficients. Therefore, the most common setup, i.e. *0_d_a*, is supposed as the optimum for AF classification achieving maximum accuracy with regard to the size of the feature vector and thus the total number of MLP parameters. These setups are used as MFCC or PLP baseline features for further analyzes realized in this work.

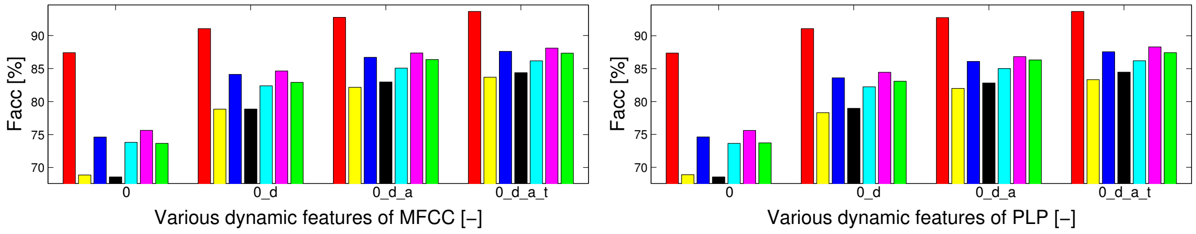


Figure 4.1: The evaluation of AF estimation for automatically labelled test set. Bars: red - *voicing*, yellow - *placed_consonant*, blue - *placed_vowel*, black - *manner_consonant*, cyan - *manner_vowel*, magenta - *rounding*, green - *sonority*.

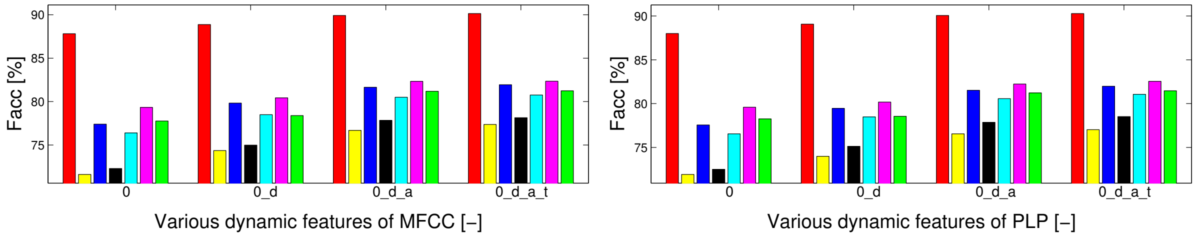


Figure 4.2: The evaluation of AF estimation for manually labelled test set. Bars: red - *voicing*, yellow - *placed_consonant*, blue - *placed_vowel*, black - *manner_consonant*, cyan - *manner_vowel*, magenta - *rounding*, green - *sonority*.

	Out	*0*				*0_d*				*0_d_a*				*0_d_a_t*			
	units	*hids*	*Ep.*	*test*	*test m.*	*hids*	*Ep.*	*test*	*test m.*	*hids*	*Ep.*	*test*	*test m.*	*hids*	*Ep.*	*test*	*test m.*
Voicing	3	700	3	87.6	87.8	400	3	91.2	89.0	1000	3	92.9	89.9	1100	4	93.8	90.1
Place_con	9	700	1	69.1	71.6	800	5	79.1	74.4	2000	5	82.2	76.5	2400	5	83.7	77.3
Place_vow	5	2200	1	74.9	77.9	500	5	84.2	79.9	1500	4	86.6	81.6	2000	4	87.6	81.7
Manner_con	9	1500	3	68.7	72.2	400	5	79.0	74.9	2400	5	82.9	77.5	700	4	84.4	77.9
Manner_vow	5	1300	1	73.9	76.7	400	3	82.4	78.6	600	4	85.0	80.5	2000	4	86.1	80.6
Rounding	4	2200	1	75.7	79.7	2400	3	84.6	80.2	1800	4	87.3	82.2	1500	4	88.0	82.1
Sonor	4	2200	1	73.8	77.6	1100	3	82.9	78.3	1500	5	86.1	80.8	2200	5	87.2	81.0
avg.				**74.8**	77.6			**83.3**	79.3			**86.1**	81.2			**87.3**	81.5

Table 4.15: Setup size of MLP for the best results with mfcc features. (Ep. = Epoch)

	Out	*0*				*0_d*				*0_d_a*				*0_d_a_t*			
	units	*hids*	*Ep.*	*test*	*test m.*	*hids*	*Ep.*	*test*	*test m.*	*hids*	*Ep.*	*test*	*test m.*	*hids*	*Ep.*	*test*	*test m.*
Voicing	3	500	3	87.6	88.0	600	3	91.2	89.1	800	3	92.9	89.9	1100	4	93.8	90.2
Place_con	9	2200	1	69.0	71.6	400	5	78.5	74.0	1800	5	82.0	76.5	2000	5	83.4	76.9
Place_vow	5	2200	1	74.9	78.0	2400	3	83.7	79.5	1300	3	86.1	81.3	1100	5	87.5	81.9
Manner_con	9	1800	1	68.6	72.4	800	5	79.1	75.1	1800	5	82.7	77.6	2200	5	84.4	78.3
Manner_vow	5	100	1	73.7	77.0	2200	3	82.2	78.5	600	4	85.0	80.5	2000	4	86.2	80.9
Rounding	4	2200	1	75.9	79.8	1800	3	84.5	80.2	2000	3	86.8	82.1	1800	4	88.2	82.3
Sonor	4	2200	1	73.9	78.1	1500	3	83.1	78.5	2200	5	86.2	81.1	900	5	87.3	81.3
avg.				**75.0**	77.8			**83.2**	79.4			**85.9**	81.3			**87.3**	81.8

Table 4.16: Setup size of MLP for the best results with plp features. (Ep. = Epoch)

With regards to particular AF classes, the significant improvement of *FAcc* can be observed in the classes describing the place and the manner of articulation for consonants. The contribution of temporal derivatives in classification of the *Place_vow*, *Manner_vow*, *Rounding*, *Sonor* classes was slightly smaller than for the category of consonants. In contrast, the smallest contribution of differential features can be seen in AF class of voicing.

The summary of the best results for the particular AF classifiers is presented in Table 4.15 and 4.16. The MFCC or PLP based input features of the MLP classifier achieved very similar results. The average accuracy across all AF features for particular temporal derivative ranged from 74.8% for *MFCC_0*, 83.3% for *MFCC_0_d*, 86.1% for *MFCC_0_d_a* to 87.3% for *MFCC_0_d_a_t*. The results for the optimum setup MFCC or PLP with *delta*, *delta - delta* coefficients were compared with the state-of-the-art results of AF classification for English [32].

II. Optimization of MLP size for AF estimation

The optimum setup of the number of neurons in the hidden layer of MLPs for all particular AF classes and various dynamic features was also analyzed. This is one of factors having a significant influence on the achieved accuracy and duration of training of MLP-based classifier. The optimum setup was empirically analyzed in the range from 10 to 2400

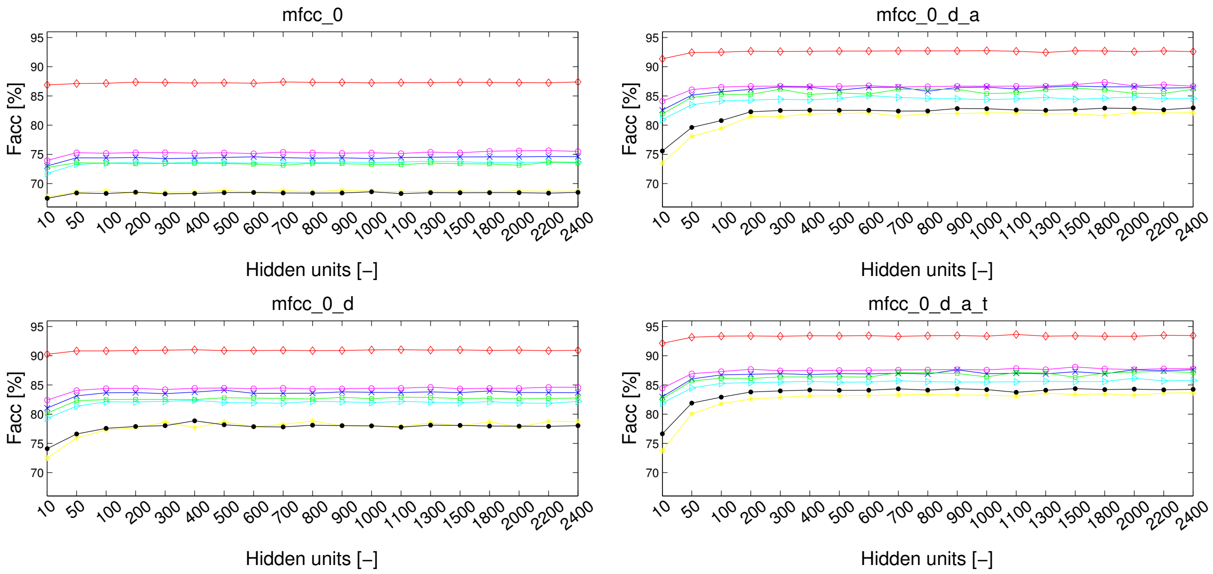


Figure 4.3: Number of hidden layer units optimization for input feature of MFCC. Lines: red ◇ - *voicing*, yellow ∗ - *placed_consonant*, blue × - *placed_vowel*, black • - *manner_consonant*, cyan ▷ - *manner_vowel*, magenta ○ - *rounding*, green - *sonor*.

hidden neurons. These analyzes are presented for the MFCC features in Fig. 4.3. Here, the dependency of target classification accuracy on the number of hidden neurons in MLP is presented. The optimum setup for the static feature vector (*MFCC_0* or *PLP_0*) is about 50 neurons in the hidden layer of MLP for all particular AF classifiers. When context information is included in the form of delta features, the optimum setup is about 200 up to 400 neurons in the hidden layer for all AF classes. The exact value depends also on cardinality of particular class classifier and target optimum setup is summarized in the following points:

- 200 for *Voicing*,
- 400 for *Place_con*,
- 300 for *Place_vow*,
- 300 for *Manner_con*,
- 300 for *Manner_vow*,
- 300 for *Rounding*,
- 300 for *Sonor*.

Conclusions

The experiments confirmed the contribution of differential features in the task of AF estimation. Both features (MFCC/PLP) achieved similar accuracy of AF classifications. The optimum setup of the number of neurons in the hidden layer of MLP for particular AF and for various acoustic inputs has been found and it will be used for the purposes of the comparison with other approaches of AF classification within the further analyzes in this work.

4.2.2 Extended temporal context in AF estimation

This section continues analyzing of AF estimation with other approaches of temporal information based on splicing short-time features at the input of neural network. Such temporal context has been proposed in hybrid MLP/HMM ASR systems where the input feature vector to the MLP classifier is composed from neighbouring feature vectors defined by a context windows of various length [91]. Generally, the purpose of using time context information in ASR is to describe the process of coarticulation produced within the human speech production. The contextual window is commonly made of MFCC/PLP and their delta, delta-delta coefficients. This approach was analyzed by other authors showing that the temporal contextual information was very important for increasing accuracy of ASR systems [92] or phone recognition[120]. The significance of contextual information in the task of phone recognition was analyzed in detail for English language in [103], [102] and it was found that the suitable length of the context should be around 90-110 ms. With regard to the usage in AF classification, the optimal length of the context window for particular AF classes was analyzed in [65]for the English language. Nevertheless, the length of 90 ms is standardly used by other authors for AF estimation [61], [120], [119].

Also further approaches of context information incorporating were proposed by other authors, e.g. DCT-TRAP, wLP-TRAPS based on long temporal context information proceeding (TRAP - TempoRAl Pattern) which was proposed by (Hermansky and Sharma 1998). The TRAP feature extraction technique is based on using the temporal trajectories of spectral power in the individual critical bands. The authors in [128], [38] showed that TRAP based features can significantly improve the performance of ASR systems and phone recognition and they have become common for front-end processing in the state-of-the-art ASR systems. Possible inclusion of a longer context at the input of ANN-based AF classifier was discussed in [137]. Since only two works [112], were found in connection with the application of TRAP to the estimation of AF, the results presented in this subpart are focused on an analysis of TRAP-based AF clasification for Czech and English.

The temporal context information is at first included using a context window created from several neighbouring short-time frames for MFCC/PLP and their dynamic coefficients and then it is obtained also using DCT-TRAP features. At the end, this section presents also the results of direct or AF-based phone recognition for both languages. Finally, both approaches are compared with MFCC/PLP baseline results described in the previous section.

Experimental setup

Regarding the parametric representation of speech signals, the context window was created from MFCC/PLP baseline features as defined above and then DCT-TRAP features were used. The setup of DCT-TRAP features (standard TRAPs with the dimension reduction using the discrete cosine transform) is summarized in the following points:

- DCT-TRAP,
  - preemphasis coefficient 0.97,
  - short-time FFT frame length of 25 ms and frame step of 10 ms,
  - 22 filters of auditory spectral analysis,
  - temporal pattern was computed from 50ms to 1s (5÷101 frames),
  - each temporal pattern was transformed to 16 DCT coefficients.

Results for temporal context represented by context window

All experiments were realized with the data from SPEECON database described in section 3.2, exactly with the OFFICE subset summarized in Table 3.2. The optimum length of context information in AF estimation was analyzed in the two similar scenarios of the experiments. Two groups of experiments were focused on finding the optimum length of context information for the case where the context window was created either from static or dynamic features.

I. Initial tuning of MLP size

The dependency of *FAcc* on the number of hidden neurons is shown in Fig. 4.4. The optimum settings across AF classes is in the range 200 ÷ 600 hidden neurons and the best setup of the MLP which achieved the best results in the estimation of AFs are summarized in Table 4.17.

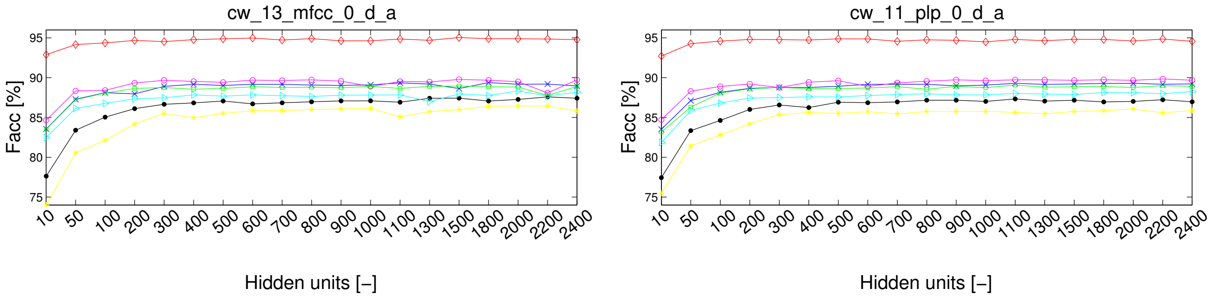


Figure 4.4: Number of hidden neurons in MLP optimization for input PLP or MFCC feature. Lines: red ◇ - *voicing*, yellow ∗ - *placed_consonant*, blue × - *placed_vowel*, black • - *manner_consonant*, cyan ▷ - *manner_vowel*, magenta ○ - *rounding*, green - *sonor*.

	Out	*MFCC_0_cw_21*			*PLP_0_cw_21*			*MFCC_0_d_a_cw_13*			*PLP_0_d_a_cw_11*		
	units	*hids*	*Epoch*	*CV*	*hids*	*Epoch*	*CV*	*hids*	*Epoch*	*CV*	*hids*	*Epoch*	*CV*
Voicing	3	700	4	94.9	2200	4	94.7	500	4	95.1	1500	4	95.3
Place_con	9	2000	8	86.2	1500	9	85.9	2000	6	86.7	2000	8	86.9
Place_vow	5	2200	5	88.6	2400	5	88.4	2000	4	89.2	1800	4	89.1
Manner_con	9	2200	8	87.1	1800	7	86.9	1100	5	87.3	2200	7	87.9
Manner_vow	5	2000	5	87.4	1300	5	87.3	2400	4	88.0	2000	5	88.1
Rounding	4	1800	5	89.2	2000	5	88.8	2200	4	89.9	1500	4	89.9
Sonor	4	1800	5	88.6	900	5	88.2	1100	5	89.2	1500	5	89.1

Table 4.17: The best setup size of MLP for the best results classification of AFs.

II. Optimization of context window length

The first experiments were focused on modelling of the contextual information based on the context window with static MFCC or PLP features only. The context window size in the range from 3 to 61 frames was analyzed and results are shown in Fig. 4.5 and the best size of the context window is marked by blue color bar. The average absolute improvement of the accuracy of AF classification depending on the varying size of the context window was compared to the zero context.

Secondly, the similar scenarios were applied also to MFCC or PLP with differential features. The results of experiments are presented also in Fig. 4.5. In this case the context window size was analyzed in the range from 5 to 31 frames and the average absolute improvement is also presented using a bar graph. The contribution of differential features brings the improvement of 0.3% in the case of MFCC and 0.9% for PLP in contrast to working with static coefficients only. This small improvement causes the input vector of higher dimensions 507 vs. 273 for the window with static parameters, moreover, the effect on the training time also increases or the data set etc. In the case of PLP the size is in the ratio of 429 vs. 273. In view of this fact, it is better to use the context window created from static MFCC or PLP features. All results presented in Fig. 4.5 were evaluated versus automatically labelled test data set.

Finally, the optimum lengths of contextual information for both types of features achieving the best results of AF classification are summarized in Table 4.18. Results for the frame level accuracy evaluated against the manually labelled test set are also presented. The best results of AF classification were achieved for PLP differential features with the context of 11 frame (plp_0_d_a_cw_11). The average accuracy across all AF features for particular best setups achieved about 89%. The *Voicing* class was classified with accuracy about 95%. The *FAcc* for the classes such as *Place_vow*, *Manner_con*, *Manner_vow Rounding*, *Sonor* achieved in range $88 \div 89\%$ and slightly worse accuracy was achieved for *Place_con* class about 87%.

To conclude this part, the detailed analysis of the optimum length of contextual infor-

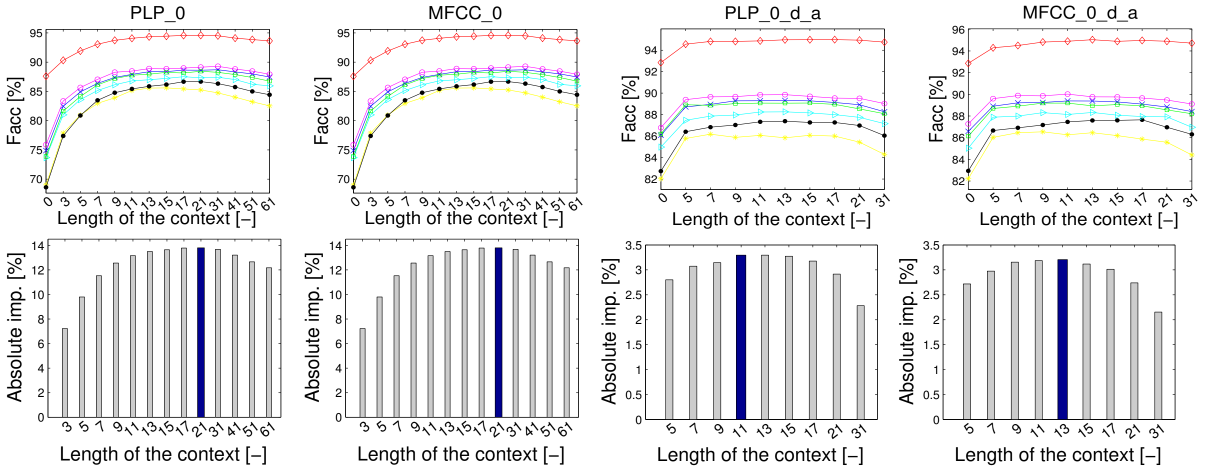


Figure 4.5: The evaluation of AF estimation for lengths of context information for static features. Lines: red ⋄ - *voicing*, yellow ∗ - *placed_consonant*, blue × - *placed_vowel*, black • - *manner_consonant*, cyan ▷ - *manner_vowel*, magenta ○ - *rounding*, green - *sonority*.

	MFCC				*PLP*			
	0		*0_d_a*		*0*		*0_d_a*	
	cw_21		*cw_13*		*cw_21*		*cw_11*	
	test	*test m.*	*test*	*test m.*	*test*	*test m.*	*test*	*test m.*
Voicing	94.7	90.6	94.9	90.6	94.6	90.7	95.0	90.7
Place_con	86.1	79.2	86.1	79.1	85.4	79.0	86.5	79.2
Place_vow	89.0	82.5	89.3	82.8	88.7	82.6	89.4	82.6
Manner_con	87.1	79.8	87.3	79.9	86.7	79.8	87.6	80.1
Manner_vow	87.5	81.5	88.2	81.8	87.5	81.9	88.3	82.0
Rounding	89.4	83.0	89.8	83.3	89.0	83.1	89.8	83.1
Sonor	88.5	81.5	89.1	82.0	88.2	81.7	88.9	81.9
avg.	**88.9**	82.5	**89.2**	82.8	**88.5**	82.6	**89.4**	82.8

Table 4.18: The best results estimation of AFs for the optimum length of context information.

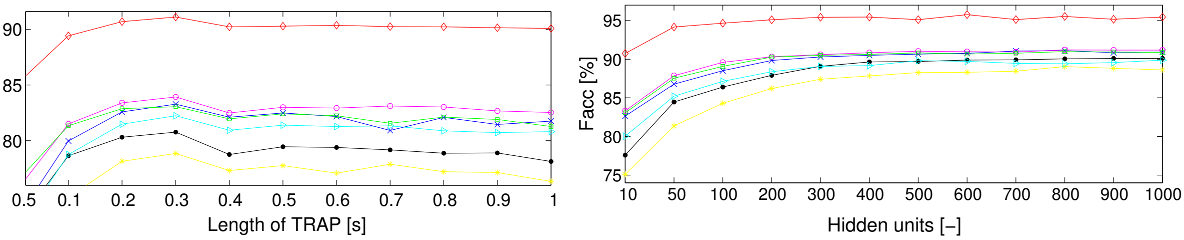


Figure 4.6: The evaluation of AF estimation for Czech; Features setup: dct-trap, 310 ms; Lines: red ⋄ - *voicing*, yellow ∗ - *placed_consonant*, blue × - *placed_vowel*, black • - *manner_consonant*, cyan ▷ - *manner_vowel*, magenta ○ - *rounding*, green - *sonority*.

mation for AF classification showed that the optimum length for the modelling of context information is between 150 and 210 ms for the static parameters. When differential features were used, it decreased to 110 ÷ 130 ms for MFCC-based features and to 90 ÷ 13 ms for PLP-based ones.

	TRAP_31_16					*TRAP_51_16*				
AF class	hid. units	CV	epoch	Test	*Test m.*	hid. units	CV	epoch	Test	*Test m.*
Rounding	900	91.2	6	89.3	*83.9*	800	89.6	5	89.2	*83.0*
Voicing	600	95.8	4	94.5	*90.5*	800	95.0	4	94.4	*90.4*
Place_con	1000	88.6	7	84.9	*78.8*	1000	85.7	5	84.0	*77.8*
Place_vow	800	91.1	7	88.8	*83.3*	1000	88.9	5	88.7	*82.7*
Manner_con	1000	90.1	8	87.0	*80.8*	800	87.4	5	86.5	*79.6*
Manner_vow	1000	89.9	7	87.5	*82.2*	900	87.5	5	87.2	*81.5*
Sonor	1000	90.9	6	88.9	*83.1*	900	89.1	5	88.9	*82.3*
avg AF	-	91.1	-	88.7	*83.2*	-	89.0	-	88.4	*82.5*

Table 4.19: The best results estimation of AFs for Czech.

Results for temporal context represented by DCT-TRAP

The following sub-part continues with experiments focusing on the analysis of a contribution of temporal context to Czech AF estimation using TRAP-based features. The main purpose is. again, to look for the optimum length of context information in DCT-TRAP features as well as to find the optimum size of the MLP in this case.

Similarly to the previous experiment, the *initial optimization of MLP size* was performed in the first step. For each AF class and each TRAP length, the optimum setup of the number of neurons in the hidden layer of MLPs was evaluated, empirical analyses were performed exactly in the range from 10 to 1000 hidden neurons and right part of Figs. 4.6 shows the results of *FAcc* for the varying size of MLP hidden layer are again in Figs. 4.6 on the left part. The optimum setup across AF classes is in the range from $400 \div 800$ hidden neurons.

Then *optimimum TRAP length* was looked for, exactly AF estimation for the lengths of TRAP-based context information in the range from 50ms to 1s was analyzed. The obtained results are presented in Fig. 4.6 and the optimum length of TRAP trajectories for AF estimation was found to be around 300 ms. Two best setups of AF estimation are summarized in Tab 4.19. The best values of *Facc* for Czech were achieved for the length of TRAP trajectories of about 310 ms for all AF classes.

The achieved results of this analysis proved that DCT-TRAP features with the given optimum length of the temporal pattern represent a suitable speech representation for AF estimation for both languages.

Partial conclusions

The best obtained results from the performed experiments are compared among themselves as well as with baseline features for Czech and they are summarized in Table 4.20. All results for techniques using some temporal context significantly surpass the results obtained by baseline features. The DCT-TRAP features seem to be very good in the task of AF classification and will be further analyzed with respect to the robust estimation under adverse conditions.

	baseline features		*context window*				*DCT-TRAP*
AF class	*mfcc*	*plp*	*cw_21_mfcc*	*cw_13_mfcc_0_d_a*	*cw_21_plp*	*cw_11_plp_0_d_a*	*dct_trap_51_16*
Voicing	92.9	92.9	94.7	94.9	94.6	95.0	94.4
Place_con	82.2	82.0	86.1	86.1	85.4	86.5	84.0
Place_vow	86.6	86.1	89.0	89.3	88.7	89.4	88.7
Manner_con	82.9	82.7	87.1	87.3	86.7	87.6	86.5
Manner_vow	85.0	85.0	87.5	88.2	87.5	88.3	87.2
Rounding	87.3	86.8	89.4	89.8	89.0	89.8	89.2
Sonor	86.1	86.2	88.5	89.1	88.2	88.9	88.9
avg AF	86.1	85.9	88.9	89.2	88.5	89.4	88.4

Table 4.20: Comparison the best results of AF classification. for Czech

4.2.3 AF estimation under adverse acoustic conditions

This part is focusing on the robust AF estimation for Czech language and the performance of the MLP classifiers under adverse acoustic conditions was analyzed. Most published works do not deal with the data gathered under adverse background conditions because the experiments in published works are usually conducted with the TIMIT database which contains speech data recorded under low noise conditions [60], [61]. Some analysis of noise robustness can be found in [65] describing the experiments with noisy data from the Verbmobil database using special MODSPEC preprocessing [63] of input features and showing that AF based ASR system works very reliably in a high noise levels environment.

Within the experimental part the basic accuracy of AF estimation for Czech using three different speech feature vectors was first tested and a detailed analysis of the optimum number of neurons in the hidden layer of MLP network was made. In the second phase, the robustness of this estimation for speech collected under various conditions from the point of view of signal quality was tested.

Experimental setup

The baseline features and DCT-TRAP features with the best setup described in the previous part were used. Speech data for these experiments were taken again from the Czech SPEECON database, the OFFICE subpart (clean speech) and CAR subpart (more noisy speech) were used, see section 3.2.

Results & Discussion

Optimization size of MLP

Firstly, the dependency of the frame accuracy on the number of hidden neurons is again presented illustratively in Fig. 4.7 for TRAP features at the MLP input and for 4 channels with different SNRs of collected speech signal. The optimum setup for particular AF classes was found to be in the range from 300 to 500 hidden neurons for *Voicing*, *Rounding*, *Sonor* and from 600 to 800 *Manner_vow*, *Place_vow*, *Manner_con*, *Place_con* across all channels.

Results for MLP-based AF classifier for which the best *FAcc* was achieved are summarized in Table 4.21 for both analyzed environments. In comparison to [65], we did not use any special preprocessing for noisy data. The same setup was used for both background conditions, i.e. for the OFFICE and CAR environment.

	Out	*OFFICE*											
	units	*Channel CS0*				*Channel CS1*				*Channel CS2*			
		hids	*CV*	*Epoch*	*test*	*hids*	*CV*	*Epoch*	*test*	*hids*	*CV*	*Epoch*	*test*
Voicing	3	2000	94.9	4	94.3	300	94.2	4	93.6	2200	92.5	4	91.6
Place_con	9	1800	85.6	5	83.9	2200	83.8	5	82.0	1500	82.3	5	80.7
Place_vow	5	1500	88.5	5	88.3	1000	87.1	5	86.9	2200	85.5	5	85.2
Manner_con	9	1800	87.2	5	86.5	1800	85.6	5	84.2	2400	84.0	5	83.0
Manner_vow	5	2400	87.0	5	86.9	1300	85.8	5	85.6	1800	84.1	5	83.7
Rounding	4	1300	89.3	5	89.1	2400	88.0	5	87.8	1000	86.5	5	86.4
Sonor	4	1300	89.0	5	88.9	2200	88.0	5	87.5	1500	86.1	5	85.8
		CAR											
Voicing	3	1000	94.6	4	87.0	1300	92.9	3	85.3	2200	91.9	4	83.9
Place_con	9	1500	85.6	8	77.8	2400	83.0	7	75.7	2000	82.0	7	74.6
Place_vow	5	2200	88.8	4	80.1	1500	87.2	5	79.2	500	86.2	5	77.7
Manner_con	9	2400	86.2	7	79.1	2000	84.0	6	77.0	1500	82.9	7	75.7
Manner_vow	5	1800	87.5	4	79.0	2200	86.3	5	78.4	2400	84.8	5	77.0
Rounding	4	2000	89.4	5	81.3	2000	87.6	5	80.4	2000	86.5	5	78.8
Sonor	4	1000	87.9	6	80.8	2000	86.2	4	78.7	2200	85.5	6	78.0

Table 4.21: Optimum setup size of MLP for the best results with DCT-TRAP features.

Robustness of MLP-based estimation of AF

The results obtained for OFFICE environment are summarized in Fig. 4.8. These results for MFCC and PLP features proved reliable standard estimation of AFs for Czech which is comparable to the results of other authors. The best results were obtained for DCT-TRAP features and for high-quality CS0 channel, i.e. 94.3 % for voicing, 83.9 % for place of consonant, 88.3 % for place of vowel , 86.5 % for manner of consonant, 86.9 % for manner of vowel, 89.1 % for rounding, and 88.9 % for sonoring. Concerning particular AF classes, the best results were obtained for voicing detection, the most difficult seemed to be the estimation of the place of articulation for consonats. When the environment is

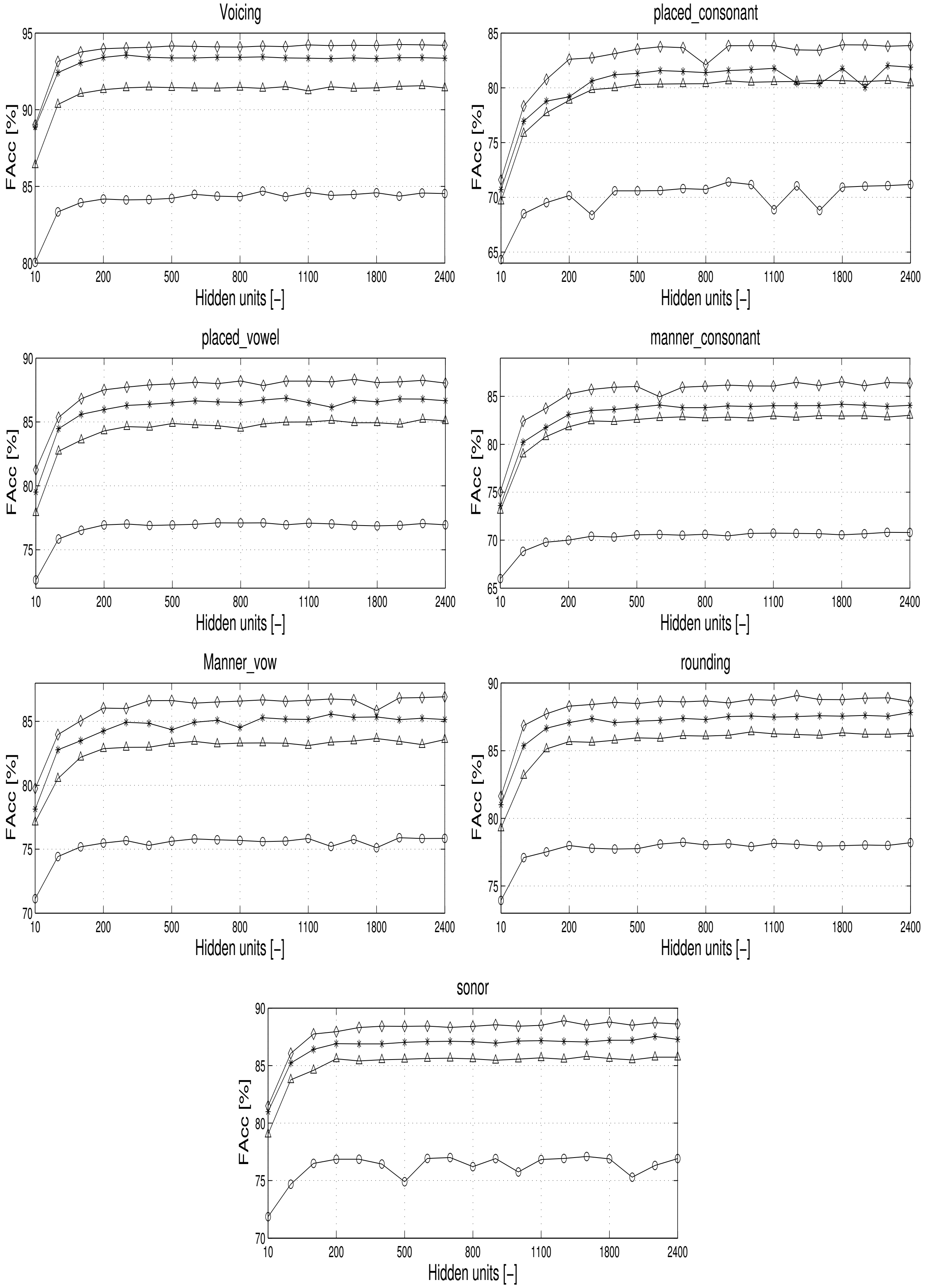


Figure 4.7: Number of hidden neurons in MLP optimization for DCT-TRAP feature; channel: ⋄ – CS0, ∗ – CS1, Δ – CS2, ○ – CS3; environment: OFFICE.

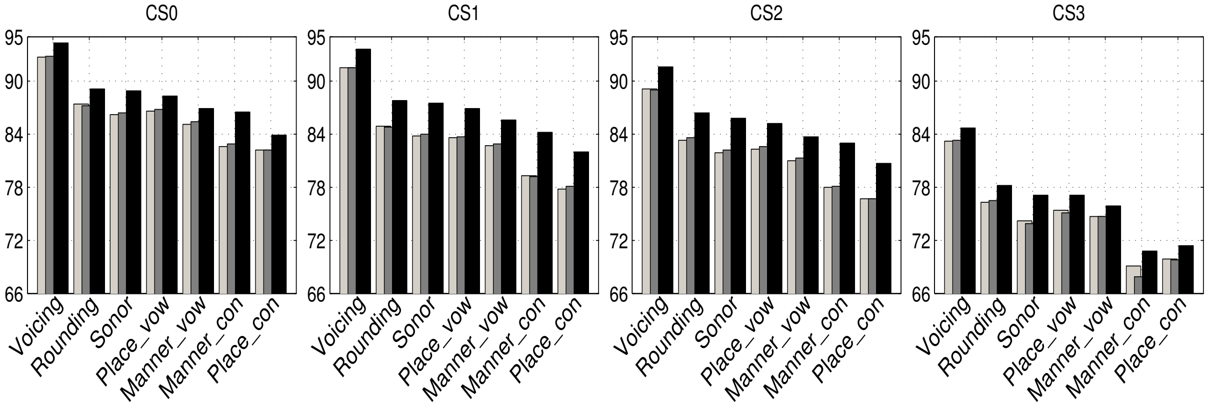


Figure 4.8: *FAcc* of AF estimation for automatically labelled OFFICE test set; features: MFCC- light gray, PLP - dark gray, DCT-TRAP - black.

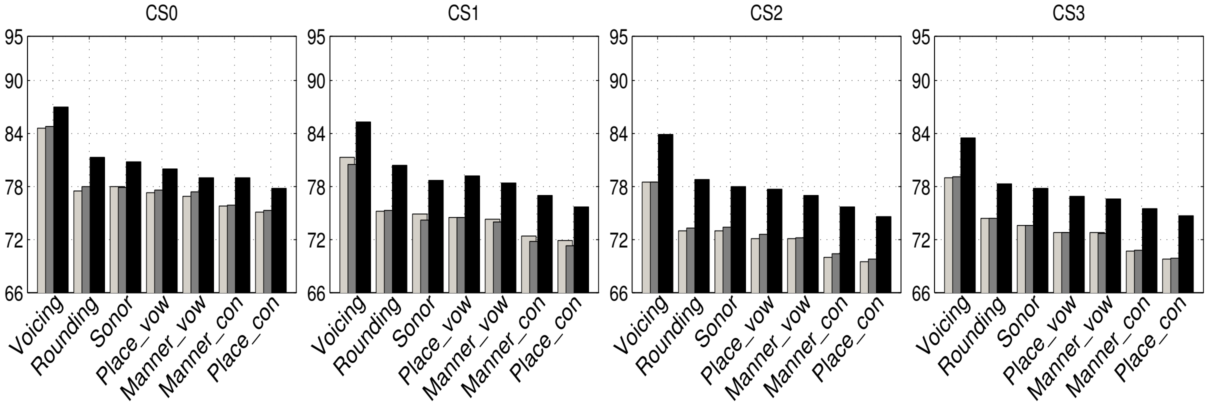


Figure 4.9: *FAcc* of AF estimation for automatically labelled CAR test set; features: MFCC- light gray, PLP - dark gray, DCT-TRAP - black.

rather clean, i.e. standard office environment, only slightly worse results were obtained for other, more noisy, channels (CS1, CS2 and CS3) with DCT-TRAP features.

Evaluations with manually labeled reference data were realized too and obtained results are in Table 4.22 using the average *FAcc* (calculated across AF). Because the evaluation with manually labeled data represented mismatched conditions, better results were always achieved in evaluations with automatically labeled reference data (automatically set boundaries were used for the training). In each case, these results have similar trend as those obtained by reference data labeled automatically.

Results obtained for more noisy CAR environment are in Fig. 4.9 and in Table 4.22 and they proved the robustness of MLP-based AF estimation, especially, when DCT-TRAP features were used as the output of acoustic analysis. We can see rather small decrease of *FAcc* in comparison to results obtained for rather clean speech data from OFFICE environment. However, for the comparison of results from these two environments, we must note that channels CS2 and CS3 contain speech of slightly different quality.

	MFCC				PLP				DCT-TRAP			
	CS0	CS1	CS2	CS3	CS0	CS1	CS2	CS3	CS0	CS1	CS2	CS3
OFFICE	81.4	79.7	78.4	73.6	81.6	79.9	78.6	73.5	82.3	81.3	80.3	74.5
CAR	85.2	83.6	81.3	81.6	85.3	83.4	81.4	81.6	85.0	85.3	83.8	83.5

Table 4.22: Average *FAcc* of AF estimation for manually labelled.

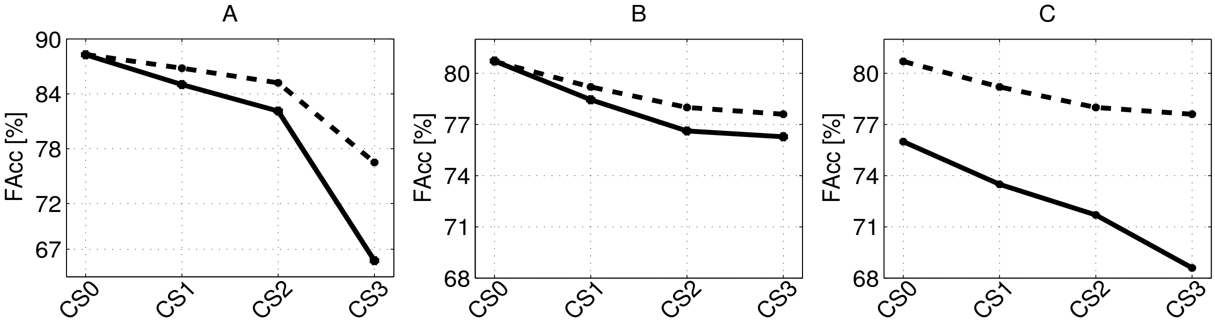


Figure 4.10: Average *FAcc* (across all AFs) for mismatched condition: A - channel mismatch in OFFICE environment (dashed line - matched training, solid line - training on CS0); B - - channel mismatch in CAR environment (dashed - matched training, solid - training on CS0); C - environmental mismatch in CAR environment (dashed - matched training in CAR; solid - training on OFFICE data).

The robustness of MLP-based AF estimation was also observed when the training and testing conditions were not same, i.e. in the case of various mismatch because it is common situation in real deployed systems having the significant influence on speech recognition accuracy [117]. These analyzes are presented using the average *FAcc* (across all AF) trend and the results for the channel mismatch and environment mismatch are summarized in Fig. 4.10. The impact of switching from close-talk microphone to the far-talk one is presented in Fig. 4.10A and 4.10B. The robustness of DCT-TRAP AF estimation is demonstrated by very small decrease of average *FAcc* when training was realized on CS0 channel only, especially in the case of CAR environment. The highest decrease was observed for CS3 channel in OFFICE environment but in this case it is the result for significantly degraded speech which SNR is typically about 6 dB only. Environmental mismatch has higher influence as it is demostrated in Fig. 4.10C. The decrease of average *FAcc* of AF estimation in CAR environment was about $6 \div 9\%$ for particular channels when training was done on OFFICE data.

Finally, the decrease of the detection accuracy of particular AFs influenced by the car noise is presented in Table 4.23. Regarding particular AFs, the decrease is about $4 \div 6\%$ and the best results were achieved for *Voicing* with average *FAcc* across all channels about 84.9%; for other AF classes it was in the range $77\% \div 80\%$ for *Rounding*, *Sonor*, *Place_vow*, and *Manner_vow*; according to generally worse *FAcc* for *Manner_con* and *Place_con* classes, the values were about 76% in this case. In the end, the realized experiments proved that

	OFFICE					*CAR*				
	CS0	*CS1*	*CS2*	*CS3*	*avg*	*CS0*	*CS1*	*CS2*	*CS3*	*avg*
Voicing	94.3	93.6	91.6	84.7	*91.1*	87.0	85.3	83.9	83.5	*84.9*
Place_con	83.9	82.0	80.7	71.4	*79.5*	77.8	75.7	74.6	74.7	*75.7*
Place_vow	88.3	86.9	85.2	77.1	*84.4*	80.0	79.2	77.7	76.9	*78.5*
Manner_con	86.5	84.2	83.0	70.8	*81.1*	79.0	77.0	75.7	75.5	*76.8*
Manner_vow	86.9	85.6	83.7	75.9	*83.0*	79.0	78.4	77.0	76.6	*77.8*
Rounding	89.1	87.8	86.4	78.2	*85.4*	81.3	80.4	78.8	78.3	*79.7*
Sonor	88.9	87.5	85.8	77.1	*84.8*	80.8	78.7	78.0	77.8	*78.8*

Table 4.23: *FAcc* of DCT-TRAP AF estimation for speech degraded by car noise.

the approach using DCT-TRAPs is generally robust for the task of AF estimation.

4.2.4 AF estimation using DNN

Estimations of AF described in previous sections were realized with shallow MLP, i.e. with only one hidden layer. As deeper DNN structures are nowadays often applied in many other applications, DNN-based AF estimation is described in this section. Experiments analyzing the estimation of English AF on TIMIT database are presented firstly, because it allows a comparison with the state-of-the-art results for AF classification task on TIMIT obtained by other authors. Mentioned AF estimation was focused on the review of various ASR features such as MFCC, PLP, FBANK, MFCC-LDA-MLLT, MFCC-FMLLR and DCT-TRAP. The review of temporal context setup was analyzed as well as the tuning of DNN hyper-parameters was performed. Described Czech AF classifier using DCT-TRAP features only was further review, similarly as it was done also in the previous section (see Table 4.20). The best AF classifier is then used within ASR, phone recognition and phonetic segmentation experiments.

Experimental setup

Concerning the English AF estimation, the experiments were performed with TIMIT subsets described in 3.10. TIMIT *CORE test set* only was used for evaluations of the particular AF classifiers. Various feature pipe-lines, which are commonly used for building of AM model were analyzed with regards to AF classification. For all cases mentioned below, the short-time frame length of 25 ms and shift of 10 ms were used as well as the selected frame was weighted by Hamming window. More details of particular feature extraction setups are given in the following points:

- MFCC
  - 30 filters (low/high 100/7940 Hz cut off) in auditory based spectral analysis,
  - 12 cepstral coefficients with the additional energy value,

- PLP
  - 23 filters (low/high 100/7940 Hz cut off) in PLP-based auditory filter bank,
  - 12 cepstral coefficients with the additional energy value,
- MFCC-LDA-MLLT
  - MFCC stacked in 11 frames window size are reduced/decorrelated by LDA/MLLT,
- MFCC-LDA-MLLT-fMLLR
  - fMLLR speaker-adapted MFCC-LDA-MLLT features ,
- FBANK
  - 40 filters (low/high 100/7940 Hz cut off) in auditory based spectral analysis,
- DCT-TRAP
  - a similar setup to 4.2.2 was used,
  - 40 filters of auditory spectral analysis,
  - each temporal pattern was transformed to 12 DCT coefficients.

Optimization of DNN-based AF estimation for English

The first part of experiments realized with TIMIT CORE test set for English were focused on the review of various features pipe-line setups, which are commonly used for training of AM. The standard feature setup such as MFCC, PLP, FBANK and DCT-TRAP was extended with MFCC-LDA-MLLT and MFCC-LDA-MLLT-fMLLR features. The speaker dependent setup of AF estimation based on the MFCC-LDA-MLLT-fMLLR feature was also analyzed. As it was described in AM section, the MFCC-LDA-MLLT-fMLLR input feature are typically stacked with context of 11 frames to create 440 dimensional feature vector for traditional tied-states DNN classifier. The same setup of context window was used and fixed for all analyzed features within the first batch of AF experiments.

The obtained results are shown in Fig. 4.11 as average *FAcc* across all AF classes. This experiment was already performed with a more advanced neural network with more hidden layers, so an impact of the number of hidden layers (blue - 1 layer, red - 2 layers, yellow - 3 layers, green - 4 layers, orange - 5 layers) as well as the impact of number of neuron per layer on accuracy of AF estimation is presented in particular graphs across all features setups. The particular hidden layers of DNN were initialized using RBM pre-training and then frame cross-entropy training was followed. We observed that RBM based pre-training helps around 0.3% with compare to random initialization of DNN weights. Therefore, the RBM pre-training was used for next all experiments.

Based on the archived results, the optimum setup across all feature setup is for DNN with two layers and 1024 neurons per layer. To conclude, the MFCC/PLP cepstral features achieved the average *FAcc* below 86%, MFCC-SPLICE_5-LDA-MLLT scored with 86% and DCT-TRAP_5/ FBANK-SPLICE_5 features overcame 86% value. The speaker dependent system achieved the best value close to 89.5%.

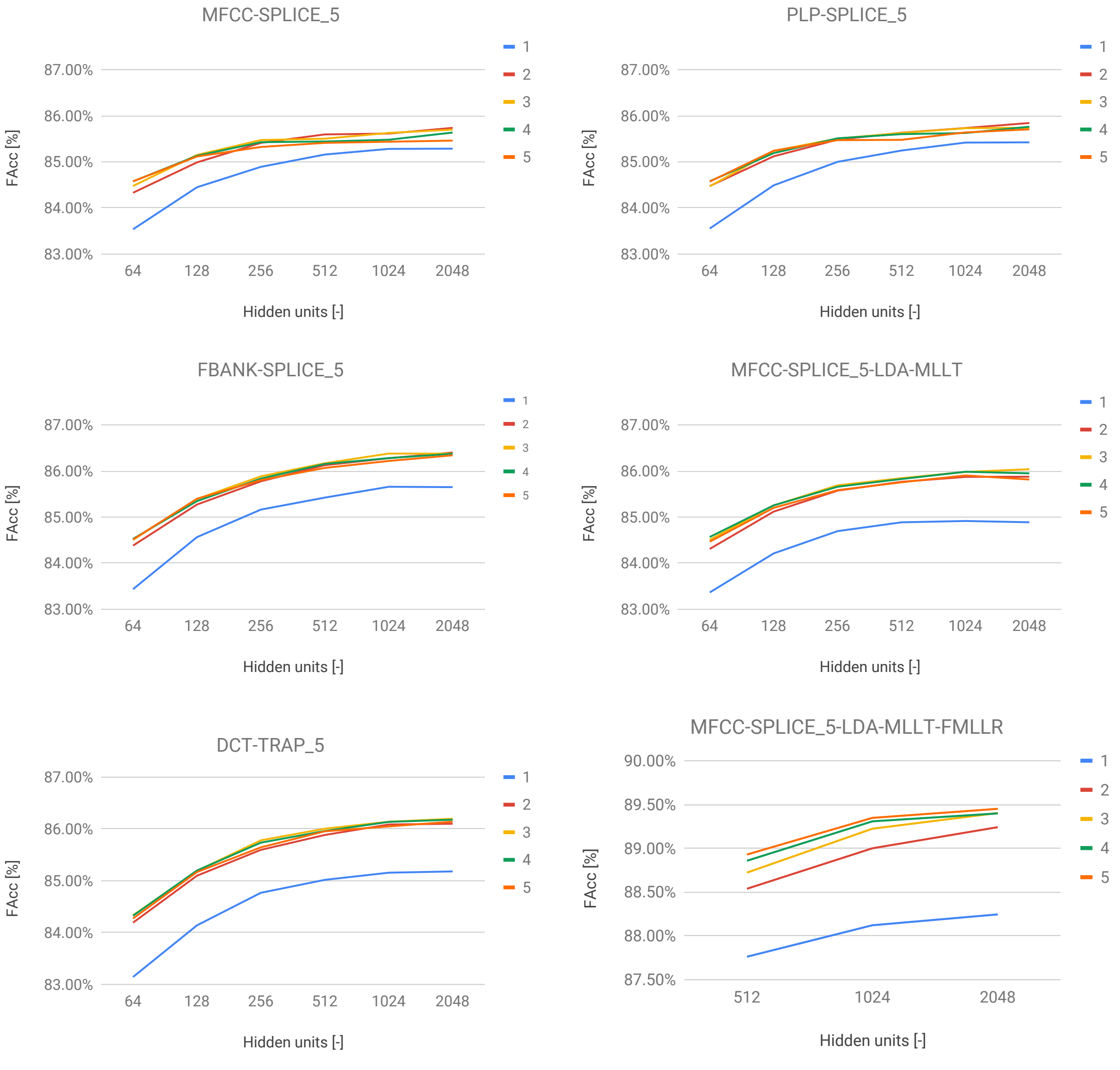


Figure 4.11: The evaluation of AF estimation for various ASR features and DNN setups.

The second group of the experiment was focused on improving speaker-independent DNN classifiers and the optimization of temporal context was analyzed. The experiments were performed with fixed DNN with two layers and 1024 neurons per layer. MFCC, PLP, FBANK and DCD-TRAP features followed same analyzes protocol which were preformed for the Czech language. The review of an optimal length of context window, was performed. The window length was analyzed in the range from 3 (0.03s) to 101 (1.01s)

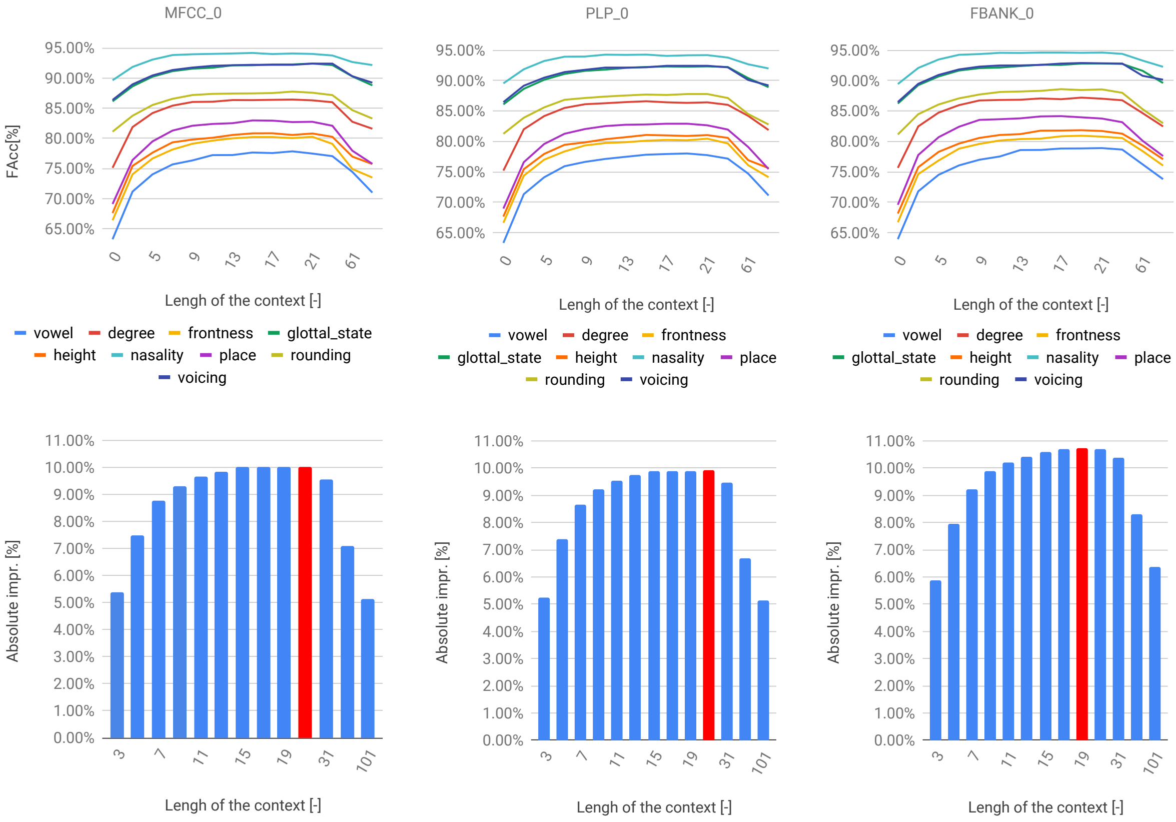


Figure 4.12: The evaluation of AF estimation for lengths of context information for static features.

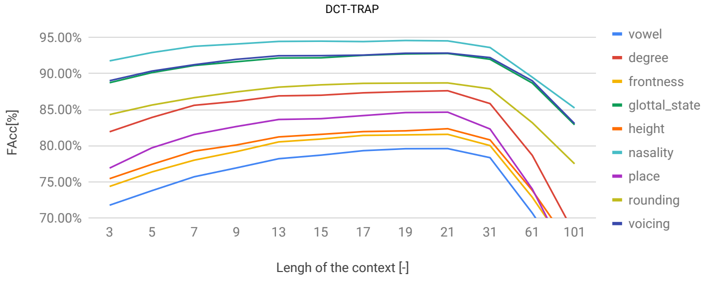


Figure 4.13: The optimization of DCT-TRAP feature.

frames and the DNN based classifier consisted from two hidden layers and 1024 neurons per layer. The results for MFCC, PLP, FBANK features are summarized in the Fig. 4.12.

The first line in the figure shows results per AF class. The second line contains the average absolute improvement of the accuracy of AF classification depending on the varying length of the window against the zero context setup (the best size is marked by red color). The optimum size of the context window length is between 19 and 21 frames for all analyzed features. The results for DCT-TRAP features are summarized in the Fig. 4.13. The optimum size of temporal pattern in DCT-TRAP setup is around 21 frames across

		AF classes									avg.
feature type	cw	vowel	degree	frontness	glottal_state	height	nasality	place	rounding	voicing	FAcc
MFCC-LDA/MLLT	5	77.53	86.63	80.36	91.32	81.04	94.25	82.59	87.54	91.61	85.87
MFCC_0	21	77.48	86.30	80.26	92.40	80.81	94.00	82.75	87.57	92.39	86.00
PLP_0	21	77.75	86.42	80.46	92.33	81.04	94.17	82.65	87.79	92.42	86.11
FBANK_0	19	78.86	87.20	80.94	92.79	81.84	94.55	83.96	88.44	92.88	86.83
DCT-TRAP	21	79.60	87.62	81.58	92.79	82.36	94.53	84.65	88.71	92.84	87.19
MFCC-LDA/MLLT -FMLLR	5	82.63	89.30	84.35	93.16	85.14	95.50	86.87	90.74	93.31	89.00

Table 4.24: The particular results for the optimized feature setup.

all AF classes.

Finally, the summary of the particular results per AF class for all analyzed features in the Table 4.24. The performed experiments confirmed the benefits of DNN based classifier and the DNN structure consisted of 2 hidden layers and 1024 hidden units was found as suitable configuration for AF classification task. In the case of results with speaker independent features, the voicing, glottal_state and nasality classes were classified with accuracy above 90%. The rounding and degree classes scored above 85% and the score around 80% achieved classes frontness, height and place. The vowel class was classified with 79%. The FBANK_0_cw_19 features overcame the cepstral features and achieved similar results as the best setup with DCT-TRAP features. With regards to optimum context window length or length of TRAP trajectories of cw_19/cw_21 frames, the results for English classes are very similar to results which were achieved for Czech language. The benefit of speaker-dependent fMMLR features was observed for AF classification task and achieved around 2% better results against speaker-dependent features.

Previously published results obtained by other authors for English and presented in [61], [120] can be summarized in the following numbers:

- *voicing*: average accuracy 90.28 %, the best 93 %,
- *place*: average 75.4 %, the best 85.9 %,
- *manner*: average 85.3 %, the best 88.5 %,
- *rounding*: average 86.21 %, the best 92 %,
- *front-back*: average 83.7 %, the best 87.4 %.

Concerning this comparison, it must be stated that these values were not always obtained under absolutely the same setup. Some authors used sometimes different AF classes, some authors also measured the accuracy at the label level, others at frame level (as it is in our case), or presented results were obtained on cross-validation set instead of test sets. Consequently, it is impossible to use a setup equivalent to all published results, but generally, it can be said that obtained results are comparable with the results obtained by other authors.

	the number of hidden layers				
	1	2	3	4	5
Voicing	94	94.28	94.4	94.42	94.4
Place_con	86.14	87.39	87.84	87.98	88.05
Place_vow	89.23	89.88	90.04	90.1	90.14
Manner_con	87.16	88.14	88.5	88.65	88.67
Manner_vow	87.95	88.72	88.88	89.07	89.1
Rounding	89.76	90.31	90.45	90.57	90.59
Sonotiry	88.98	89.67	89.97	90.05	90.11
avg.	89.03	89.77	90.01	90.12	90.15

Table 4.25: The impact of hidden layers on DNN-based AF estimation for Czech.

DNN-based AF estimation for Czech

The second part of experiments were focused on the review of Czech AF estimation in which more deeper DNN classifier was involved. DNN-based AF classifier based on DCT-TRAP features only was here evaluated. The impact of more hidden layers was analyzed on the Speecon Office CS0 test set and it is described in Table 3.2. The results are summarized in the Table 4.25. The addition of the second and the third hidden layer helps to improve average accuracy on the level of *FAcc* around 0.7% for DNN with 2 hidden layers and around 1% for DNN with 3 hidden layers.

Partial conclusions

The performed experiments for English language proved that FBANK and TRAP features represent a suitable speech representation for AF classification task. The optimum value for temporal context was found around 21 frames. The speaker depended FMLLR features improved the AF results, however, our generic goal is to build speaker-independent AF classifier. The optimum configuration of DNN based AF classifier was found for DNN structure consisted from 2 hidden layers and 1024 hidden neurons per layer.

4.2.5 AF estimation for Speechdat-E languages

Within the experiments for Czech and English languages, the analyzes of various type of features, temporal context setups and DNN structure were released mainly on 16kHz corpora. In this section, we continue with the next AF experiments with focus on East European languages from SpeechDat-E corpora which were recorded through telephone line. Finally, Russian, Slovak, Polish, Hungarian as well as Czech languages were analyzed.

Experimental setup

As it was mentioned above, the experiments were performed with E-Speechdat corpus and particular data sets were described in section 3.2 or in more details in Table 3.8.

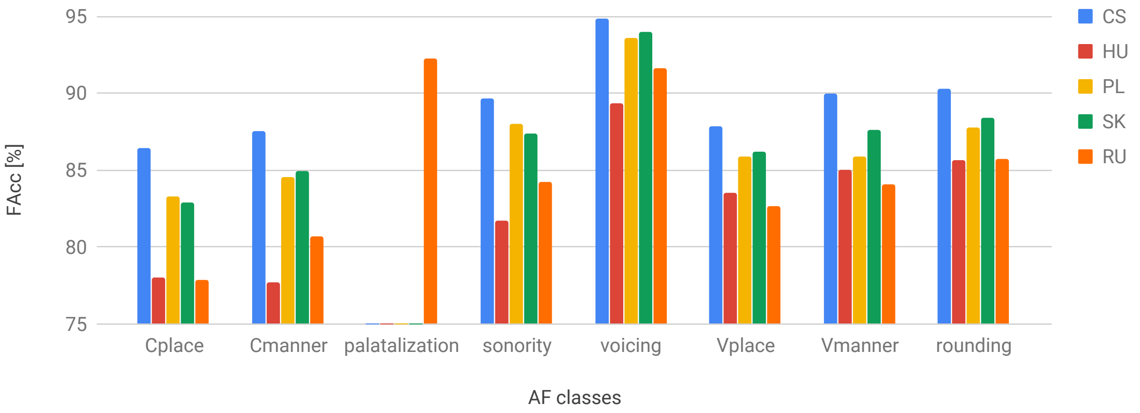


Figure 4.14: The AF results for all languages.

The DNN based AF classifier was used for these experiments, targets were created using GMM-HMM system with *tri2* AM trained per particular language, DNN was then trained using cross-entropy training, and features were computed with the following setup:

- DCT-TRAP
    - preemphasis coefficient 0.97,
    - short-time FFT frame length of 25 ms and frame step of 10 ms,
    - 20 filters (low/high 100/3800 Hz cut off) of auditory spectral analysis,
    - temporal pattern was computed from 50ms to 1s (5÷101 frames),
    - each temporal pattern was transformed to 16 DCT coefficients.

Results & Discussion

The released experiments are summarized in the Fig. 4.14 and in the Table 4.26. The best results were achieved for the Czech language which significantly overcame other languages with the average *FAcc* across all classes around 89.51%. The Polish and Slovak languages achieved average *FAcc* around 87% and the avg. *FAcc* around 83% was achieved for Hungarian and Russian languages. The *Voicing* class was classified with accuracy about 94%. The *FAcc* for the classes such as *Rounding*, *Manner_vow* and *Sonotiry* achieved *FAcc* around 90% and slightly worse *FAcc* was achieved for *Place_con, Manner_con, Place_vow* classes around 87% for the Czech. With regards to AF class *FAcc* across languages, the *Voicing* class was classified with accuracy 92%, the classes *Rounding, Manner_vow, Place_vow, Sonotiry* achieved *FAcc* around 85 ÷ 87% and the classes *Place_con, Manner_con* were classified with *FAcc* around 81 ÷ 83%. The *Palatalization* class included in Russian language was classified with accuracy above 92%.

Language	Place_con	Manner_con	Palatalization	Sonority	Voicing	Place_vow	Manner_vow	Rounding	avg.
CS	86.41	87.5	-	89.69	94.86	87.83	89.99	90.32	89.51
HU	77.99	77.68	-	81.71	89.36	83.54	85.02	85.67	82.99
PL	83.26	84.57	-	87.97	93.62	85.91	85.85	87.76	86.99
SK	82.86	84.92	-	87.41	93.97	86.18	87.65	88.41	87.34
RU	77.85	80.7	92.27	84.2	91.65	82.67	84.08	85.74	84.90
avg.	81.67	83.074	-	86.20	92.692	85.226	86.518	87.58	-

Table 4.26: The particular AF results.

To conclude this part, it can be stated that achieved results for 5 East-European languages are comparable with achieved results for Czech and English in the previous sections. The reduction of signal spectral content given by telephone-band environment also did not have significant negative impact on estimation of AF classes, e.g. for both 8kHz and 16kHz variants of Czech AF estimation, equivalent results were achieved with average *FAcc* around 89%.

Chapter 5

Applying Articulatory Features within Speech Processing

The previous chapter described the design of AF classifier and presented the performance of implemented classifiers on the level of *FAcc*. In this chapter, a number of possible applications of AF are presented. First, a tool for pure visualization of estimated AF classes around a given speech signal is presented. Further, the integration of AF features into acoustic modelling of speech recognizer is described as a key application of AF. It is implemented in the form of AF-based TANDEM system. The experimental part related to his task analyzes the contribution of AF features to phone recognition as well as to ASR under various acoustic conditions tasks. Further supposed ASR application is in phonetic segmentation algorithm, however, it is described separately in next chapter. This chapter ends by the discussion about a potential usage of the AF classifiers within automatic clinical assessment of pathological speech disorder.

5.1 Visualization of estimated AF

An analyzing of AF estimation is realized typically on selected testing datasets because such results can give statistically significant classification of AF estimation accuracy. On the other hand, to understand better the behaviour of AF classifiers, an observation of obtained results aligned around selected particular utterance is very illustrative and useful. The information about articulations obtained on the basis of AF can benefit also general phonetic research as well as the manual phonetic segmentation which must be frequently prepared for evaluation sets for phonetic segmentation or some other basic ASR tasks.

This visualization was realized in Praat-tool [10] and one illustrative example of TextGrid with information about estimated AF is shown in Fig. 5.1. We can see here a signal and its spectrogram (potentially with other estimated features, here, pitch is

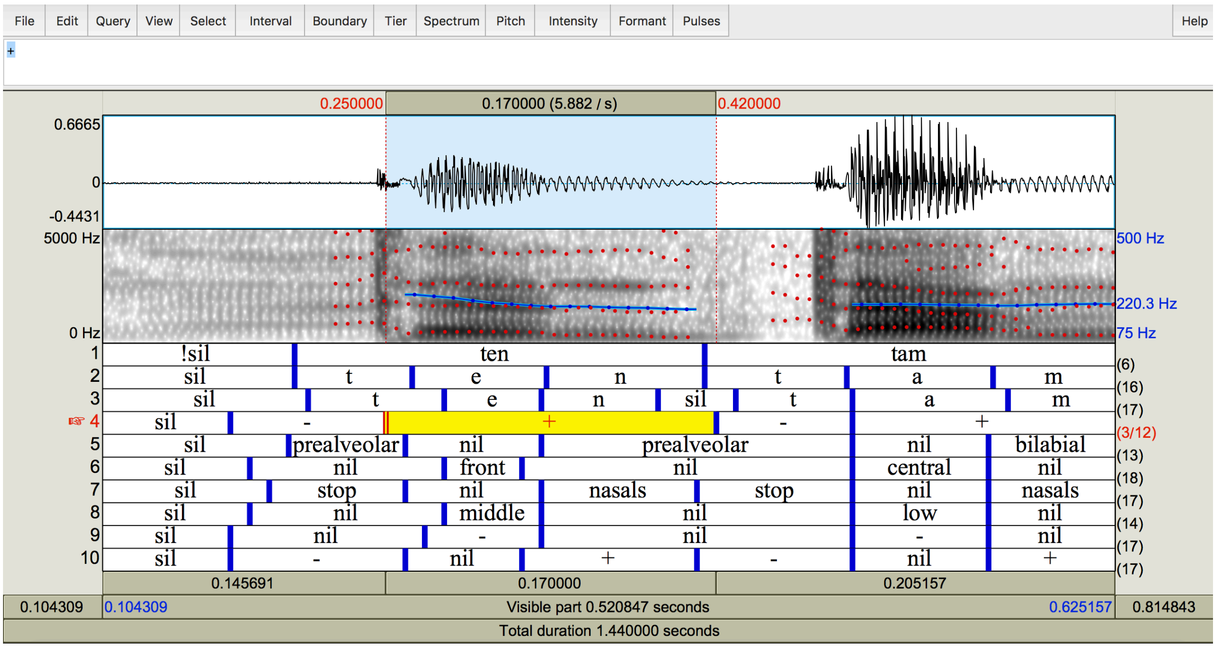

File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
0.250000
0.170000 (5.882 / s)
0.420000
0.6665
0
-0.4431
5000 Hz
0 Hz
500 Hz
220.3 Hz
75 Hz
1 !sil ten tam (6)
2 sil t e n t a m (16)
3 sil t e n sil t a m (17)
4 sil - + - + (3/12)
5 sil prealveolar nil prealveolar nil bilabial (13)
6 sil nil front nil central nil (18)
7 sil stop nil nasals stop nil nasals (17)
8 sil nil middle nil low nil (14)
9 sil nil - nil - nil (17)
10 sil - nil + - nil + (17)
0.145691
0.170000
0.205157
0.104309
0.104309
Visible part 0.520847 seconds
0.625157
0.814843
Total duration 1.440000 seconds

Figure 5.1: Illustrative example of visualized AF features.

estimated) which are aligned with the TextGrid contain 10 layers with AF information. Particular layers are defined as follows:

1 . layer: manually set word boundaries,

2 . layer: manually set phone boundaries,

3 . layer: automatically aligned phone boundaries,

4 . layer: estimated voicing,

5 . layer: estimated place of articulation for consonants,

6 . layer: estimated place of articulation for vowel,

7 . layer: estimated manner of articulation for consonants,

8 . layer: estimated manner of articulation for vowel,

9 . layer: estimated rounding,

10 . layer: estimated sonority.

The procedure of creating TextGrids with estimated AF was integrated into the implemented automatic phonetic segmentation script. Results of phonetic segmentation as well as of AF estimation is obtained in the form of very readable ctm-format. It is a text format of segmented data including time marks and illustrative example of ctm-file for *voicing* AF class is shown in next lines.

```
18_241108_s1_0261_wav 1 0.00 0.17 sil
18_241108_s1_0261_wav 1 0.17 0.08 -
18_241108_s1_0261_wav 1 0.25 0.17 +
```

Then python classes for manipulation with Praat TextGrids created by Gorman were used[1]. The tools for conversion between TextGrid and ctm-file (and vice verse) were created using above mentioned python library, as well as some tools for manipulating with particular layer in given TextGrid. Exactly we have now available the following tools:

- ctm2textgrid - it allows to load various number of ctm files are concatenated them to tiers in textgrids
- textgrid2ctm - it allows to split textgrid to specific ctm file per tier

Presented extended TextGrids with information about articulation are used in our laboratory and they help with research focus on the study of irregular pronunciations on NCCCz corpus.

5.2 AF application for ASR & Phone Recognition

As it was mentioned in section 2.3.3, AF were successfully integrated in various speech tasks. With regards to the acoustic modelling in speech recognition, the integration of the outputs from the MLP based AF classifiers into both the hybrid ANN-HMM architecture and the TANDEM architectures was analyzed by authors in [66], [85], [31], [33], [78] for various languages and the comparison of the both approaches was investigated in [78] with the conclusion that the TANDEM system achieved better results in comparison to the hybrid system based on AF. The TANDEM system was investigated in the cross-lingual ASR in [14], [73]. Recently, a joint estimation approach where AF and AM are jointly estimated was proposed in [1] and successfully applied on the low-resource languages ASR task.

In this thesis, the AF-based TANDEM approach is investigated with regards to analyze the contribution of AF in the large-vocabulary tasks. Authors in [66], [33], [14], [78] presented the results for the TANDEM architecture based on the monophone or triphone AMs, which were based on MFCC/PLP cepstral features and targeted to a rather small-vocabulary ASR task.

The AF-based TANDEM ASR system

The standard TANDEM architecture was described in section 2.2.2. The TANDEM system consists of two components, the first part with ANN classifier which extracts phone posteriors features and the second part with GMM-HMM model. In the case of the

[1] http://github.com/kylebgorman/textgrid.py

AF-based TANDEM system, the first part contains for each AF class specific ANN classifier, which produce posteriors features per AF class. These AF posteriors features are combined to final AF high-dimension vector and the logarithm operation and PCA are applied to obtain suitable feature vector for the GMM-HMM AM. Finally, the processed AF posteriors feature vector is typically concatenated with a cepstral MFCC/PLP feature vector.

5.2.1 AF-Based Phone Recognizer for English

The initial experiments focusing on the incorporation of AF into ASR system for Phone recognition task were realized. The phone recognition system was implemented based on Kaldi (*timit/s5 recipe* [2]). The standard Kaldi TIMIT recipe presents the performance of the Phone recognition on the *CORE test set* for various complexity of AMs. Kaldi reference results were used as baseline results for comparison with the AF-TANDEM system in the experimental part.

Experimental setup

As it was mentioned, the experiments in this section were performed with the TIMIT data sets, which were described in Table 3.10. The performance of the phone recognition was measured on the level of *PER* and AF-TANDEM system consisted of the AF classifiers and GMM-HMM model. The bigram LM was trained on train set with phonetic transcription and specific lexicon with pure list of phones instead of words (phone-to-phone mapping. e.g.: ah -> ah) was used for decoding purposes. The setup of articulatory classifiers is summarized in the following paragraph.

The detailed review of articulatory classifiers was presented in the previous chapter and the classifiers with the best performance on the level of *FAcc* accuracy were selected for the AF posterior feature extraction task. The 9 independent DNN based classifiers for particular AF classes such as *degree* (87.62% *FAcc*), *frontness* (81.58% *FAcc*), *glottal state* (92.79% *FAcc*), *height* (82.36% *FAcc*), *nasality* (94.53% *FAcc*), *place* (84.65% *FAcc*), *rounding* (88.71% *FAcc*), *vowel* (79.60% *FAcc*) and *voicing* (92.84% *FAcc*) were selected. The particular DNN classifier consist of 2 hidden layers and 1024 hidden units. The classifiers were trained on top of DCT-TRAP features where the length of the TRAP trajectories was 21 frames.

During the next step, the frame posterior features were extracted from particular AF classifier and concatenated to final high-dimensional vector. The dimension of AF posterior vector is 67 for English. Then, the logarithm operation is applied on AF posterior

[2] https://github.com/kaldi-asr/kaldi/tree/master/egs/timit/s5

vector and finally, the vector is de-correlated and reduced by PCA. The processed AF feature vector is concatenated with various type of feature pipe-lines depending on the type of AM complexity. In the case of the TANDEM *mono* and *tri1* systems, the processed AF feature vector is concatenated with the MFCC cepstral features and their Δ and $\Delta\Delta$ coefficients. The TANDEM *tri2* system is based on concatenated the processed AF features with MFCC-SPLICE5-LDA-MLLT features. The speaker-dependent TANDEM *tri3* system uses the processed AF features concatenated with MFCC-SPLICE5-LDA-MLLT-fMLLR features. To conclude, the experiment with hybrid DNN-HMM system was trained on concatenated speaker-dependent MFCC-SPLICE5-LDA-MLLT-fMLLR features with the processed AF feature vector.

Results & Discussion

The experiment part can be divided into two parts. The first part compares the Kaldi baseline models with the AF-Based TANDEM systems. Then, AF class specific based TANDEM system were trained to present the contribution of particular classes on the *PER*. Finally, the second part analyzes the combination of both systems together.

The results for Kaldi baseline models and implemented AF-TANDEM systems are summarized in the Table 5.2. The AF-TANDEM system based on combination of all AF classes achieved better results for *mono* and *tri1* based AM systems. The achieved results are correlated with the results of other authors in [78], which presented the same positive impact of AF-TANDEM for *mono* and *tri1* systems. The AF-Based TANDEM systems achieved significantly worse results for *tri2*, *tri3*, *dnn*, which are based on stacked feature vectors which describe temporal context and possible cover the complementary information of AF features, which helped *mono* and *tri1* system. To better understand which AF class help to reduce *WER*, the AF-TANDEM system was trained on the concatenated MFCC features and selected AF class posterior feature vector. The obtained results are summarized in the Table 5.2. The AF class *voicing* as well as *degree*, *frontness* and *glottal_state* have major impact on adding complementary information to AM trained on MFCC features. The minor contribution was observed for the classes *height nasality place* and class *vowel.*

	mono	tri1	tri2	tri3	dnn
baseline	32.7	25.6	23.7	21.6	18.5
af-tandem	27.2	24.8	25.1	27.1	19.9

Table 5.1: Achieved PER for baseline models and AF-TANDEM systems.

The next part of experiments was focus on combination of ASR systems trained with and without AF information and the results are summarized in the Table 5.3. The ASR systems combination was realized based on the lattice-level MinimumBayes Risk (MBR)

DNN-HMM	AF-DNN-HMM								
senones	all	vowel	place	nasality	heigth	glottal_state	frontness	degree	voicing
18.5	19.9	20.3	19.6	19.2	19.1	18.8	18.7	18.7	18.6

Table 5.2: Achieved PER for AF class specific DNN-HMM systems.

system combination approach [136]. The positive impact of combined AM systems was observed for *mono*, *tri1*, *tri2* and *dnn* types. A little bit worse result was obtained for *tri3* system. The big *PER* reduction was observed for *mono* system and around 1.5% for combination of DNN-HMM systems.

	mono	tri1	tri2	tri3	dnn
baseline	32.7	25.6	23.7	21.6	18.5
af-tandem	27.2	24.8	25.1	27.1	19.9
combined system	24.9	23.5	23.3	23.9	17.1

Table 5.3: The impact of AMs combination.

5.2.2 The AF-based TANDEM system for Czech

In this section, we continue with experiments focused on the Czech language. The first part analyzes the behaviours of AF-based TANDEM system in LVCSR task on SPEECON and NCCCz databases. The first experiments were realized in the close-talk channel microphone in clean office environment. Then, the behaviour of the TANDEM system under other acoustic conditions such as CAR was analyzed. Finally, the behaviour of AF-TANDEM system was reviewed on the NCCCz test set with causal speech utterances.

Experimental setup

As it was mentioned, the experiments in this section were performed with the SPEECON and NCCCz data sets, which were described in Table 3.1 and in Table 3.6. Similarly to English, the Czech AF classifiers were selected based on the review in the previous chapter. In total, 7 independent DNN based classifiers for particular AF classes such as *Place_con* (87.39% *FAcc*), *Manner_con* (88.14% *FAcc*), *Place_vow* (89.88% *FAcc*), *Manner_vow* (88.72% *FAcc*), *Rounding* (90.31% *FAcc*), *Sonority* (89.67% *FAcc*), and *Voicing* (94.28% *FAcc*) were selected. The particular AF DNN based classifiers consist of 2 hidden layers and 1024 hidden units. The classifiers were trained on top of DCT-TRAP features where the length of the TRAP trajectories was 21 frames.

The whole process of recognition with AF-TANDEM system starts with the extraction of AF posteriors from trained DNN classifiers and computation of MFCC features. The extracted posteriors features per AF class are concatenated to final 39 dimensional AF feature vector and then post-processed with logarithm operation and PCA transformation to final feature vector with 38 dimension. The processed AF feature vector is appended with ASR features in feature pipe-lines of particular AM systems such as *mono*, *tri1*, *tri2*, *tri3* and *dnn*. The tri-gram based LM with Witten-Bell smoothing was trained on the text file from train part of NCCCz corpus.

Results & Discussion

The first results describe the behaviour of AF-TANDEM system on recognition of Czech read speech. The GMM-HMM and AF-TANDEM systems were trained on the Speecon OFFICE environment and close-talk CS0 microphone. The performance of both architecture was analyzed on the matched and the mismatched acoustic conditions. In the case of the matched setup, the AMs were tested on OFFICE CS0 close-talk channel. As mismatched environment was selected CAR CS0 channel environment which contain higher amount of car noise with compare to OFFICE environment. The obtained results for matched setup of experiment are summarized in the Table 5.4. The AF-TANDEM system achieved better results on the level of *mono* and *tri1* systems. Worse results are observed for *tri2* and *tri3* system. The accuracy of *mono* and *tri1* system can be further improved using MBR system combination. The next improvements were not observed for *tri2* and *tri3* system.

AM type	GMM-HMM	AF-TANDEM	Combined
mono	21.32	20.69	19.08
tri1	15.26	17.93	14.96
tri2	15.99	19.81	16.23
tri3	15.87	19.21	17.18

Table 5.4: *WER* results on Speecon - OFFICE CS0 test set

The second part of experiments was focus on the performance of AF-TANDEM system on casual speech recognition task on NCCCz corpus. The achieved results are summarized in the Table 5.5. The achieved results shown significant better performance of *mono* AF-TANDEM system with compare to GMM-HMM architecture. Slightly smaller improvement was observed for *tri1* AF-TANDEM system. The *tri2* and *tri3* AF-TANDEM systems achived worse results with compare to GMM-HMM version. In the case of ASR system combination, the positive impact was observed for *mono* and *tri1* system. The ASR combination of *tri1* and *tri2* did not archived better results over baseline GMM-HMM system.

In this section, the AF-Based TANDEM system was presented and analyzed in the phone recognition for English and for Czech ASR tasks. The AF-TANDEM system based on *mono*, *tri1* achieved better results in comparison with standard mono and *tri1* GMM-HMM systems in both recognition tasks. The presented results confirm the results of

AM type	GMM-HMM	AF-TANDEM	Combined
mono	78.28	74.47	74.15
tri1	61.53	61.33	60.54
tri2	59.14	62.44	60.44
tri3	54.66	61.78	54.69

Table 5.5: *WER* results on NCCCz test set

other authors that the AF-TANDEM based *mono*, *tri1* systems help on small vocabulary recognition tasks [78], [33]. The benefit of the AF-TANDEM system for *tri2*, *tri3* and dnn was not observed. However, the positive contribution was observed from combination of decoded hypotheses from both systems and significant improvement *PER* reductions for *mono*, *tri1*, *tri2* and dnn AM systems was obtained. In the case of ASR experiments on Speecon and NCCCz corpora, the benefit from ASR systems combination was observed for *mono* and *tri1*. The realized experiment confirmed the usage of the developed AF classifiers for phone recognition and ASR tasks.

5.3 AF in Biomedical Speech Applications

The two previous sections presented the incorporation of AF classifiers into ASR systems and proved the quality of the trained DNN based classifiers. As other application of developed AF classifiers could be an integration into systems used in clinical biomedical applications studying particular diseases affecting voice production. This is really up-to-date topic from commercial point-of-view.

The recent research from this area has shown that about 7.6% of adults in the United states are affected by a voice disorder annually and only a smaller part of adults find professional help [9]. However, to help this smaller group of adults who ask for the health-care system, large financial and human resources are required. The automatic clinical assessment of voice disorders represent active research topic in the past years.

The mainstream approaches are based on the combination of a machine learning and signal processing tools [42]. The challenges for developing of real assistant systems, which could be used in a clinical practice are that the amount of medical speech data is very limited. Typically, collected speech data are under strict policies or the amount of speech is too limited. The other issue is that the labels have to be created by specialists and strongly depend on subjective assessment. That is why the researches are trying to transfer knowledge and frameworks from the ASR field where data is abundant [101] and systems are well developed [108], [2].

The knowledge from an ASR domain can be transferred to medical speech domain using pre-training of a system on the non-pathological speech and then fine-tune the system on the pathological speech [144]. The next approach is to train a system on healthy speakers from one database, perform the evaluation on healthy speakers and pathological speakers from another database, and compare differences. Typically, these works report acceptable results wiht regards to correlation between features and clinical intervention. The MFCC/PLP features are generally used, which makes the results difficult to understand for specialist from clinical domain. This feature makes the system unusable for deployment to real practice.

Therefore, there have been attempts to use features which are easily interpretable for these specialist. The AF feature represent the suitable features for this task due to their correlation with vocal tract properties. This approach was presented in a study [56] where the author trained multi-label RNN to map MFCCs to AF and the for training was used TIMIT database. Then, the authors used the trained model to extract AF for dysarthric and normal speakers and checked differences. The achieved results shown that AF were significantly different for dysarthric patients with compare to healthy patients. The approach was verified in a study [123] which showed that appended AF vector with standard MFCC vector can increase ASR performance for disarthric speakers. This conclusion was confirmed by [26].

Chapter 6

Phonetic Segmentation and Pronunciation Detection

This chapter deals with phonetic segmentation and pronunciation detection tasks which are applied on both the formal and casual speech. In the beginning, the performed study focused on the analysis of accuracy for various acoustic modelling techniques such as GMM-HMM vs. DNN-HMM, monophone vs. triphone, or speaker independent vs. speaker dependent. The impact of pronunciation lexicon on the accuracy of phonetic segmentation is also analyzed. These analyses are presented for English and Czech languages and a possible contribution of AF is also analyzed at the end.

6.1 Phonetic segmentation state-of-the-art

Automated phonetic segmentation is a task which has possible applications in a variety of speech technology systems. It is a procedure which defines boundary locations of particular phones in a given utterance and whose usage is necessary in situations when phone boundaries must be found for very huge corpora. It is typically used to create sub-word units for the purpose of concatenative speech synthesis [83], [121], to determine phone boundaries in a huge speech corpora for the training of neural-networks-based recognizers, for voice activity detection systems and articulatory feature classifiers, or in other applications motivated by a study of pronunciation variability based on phonetic segmentation. When phone boundaries are known, a detailed analysis of particular phone realizations can contribute to the clinical diagnostics of serious diseases which influence speech production [98] or to an analysis of pronunciation variability in a spontaneous or informal speech [89].

6.2 Phonetic Segmentation Framework

Automatic phonetic segmentation can be implemented in various ways. HMM-based automatic phonetic segmentation, which is well-known as a forced alignment, is the most widely applied technique. However, other approaches for the phoneme localization using Bayesian change-point detector, or artificial neural networks are also used by some authors [19].

HMM-based forced-alignment

The HMM-based forced alignment is a well known and basic solution which is based on the alignment of trained HMM models to a given utterance when the utterance content is known. This algorithm is based on looking for the maximum likelihood path through a decoding graph composed of an acoustic model and a grammar representing the utterance content. For this purpose, the selection of proper pronunciation, ideally the one that has been observed, plays a significant role in the segmentation accuracy. Phone boundaries are then determined by the occupancy of HMM states representing particular phones over the found optimum path. This procedure is commonly used during the acoustic model training. As it was mentioned, Kaldi toolkit was used for all experiment within this thesis but it is not very common to use deep AMs for speech segmentation [82]. It is more common to use a low-level AM which are also used to generate targets for DNN-HMM training. The task presented in this chapter makes use of speaker-independent/dependent GMM-HMM models and DNN-HMM models that were used in the previous chapters. The abbreviations for these models follow the ones set in previous chapters: *mono*, *tri1*, *tri2*, *tri3*, *dnn*.

Evaluation of phonetic segmentation accuracy

The evaluation of phonetic segmentation accuracy was done using the criteria describing both the accuracy at the level of phone recognition correctness as well as the accuracy of phone boundary placement (this approach was also used by authors in [58] or [82]). The phone recognition correctness evaluates the standard *Phone Error Rate* (defined using equation in 3.3) computed on the basis of Levenshtein distance. For the purpose of phonetic segmentation classification, it is also suitable to use *Phone Correctness* computed as

$$PCorr = \frac{N - S - D}{N} \cdot 100, \tag{6.1}$$

where N is the number of phones in the reference and S and D are numbers of substitutions and deletions in the aligned data. This criterion is more suitable because the evaluation of the accuracy of a particular boundary placement makes sense just for correctly recognized phones not for possibly inserted phones. All deleted phones are also removed from the reference transcript, inserted phones from aligned transcript, and substituted phones are removed from both of them. The cleared transcripts are then used for the evaluation of boundary placement accuracy. When two couples of reference and transcribed boundaries for each phone realization are available, i.e. $beg_{ph,ref}[i]$ and $end_{ph,ref}[i]$ vs. $beg_{ph}[i]$ and $end_{ph}[i]$, *Phone Beginning Error* (*PBE*) and *Phone End Error* (*PEE*) can be defined for each particular phone *ph* as

$$PBE_{ph}[i] = |\, beg_{ph}[i] - beg_{ph,ref}[i]\,| \,, \tag{6.2}$$

$$PEE_{ph}[i] = |\, end_{ph}[i] - end_{ph,ref}[i]\,| \,. \tag{6.3}$$

as well as *Phone Length Error* (*PLE*)

$$PLE_{ph}[i] = |\, end_{ph}[i] - beg_{ph,ref}[i] - end_{ph,ref}[i] + beg_{ph,ref}[i]\,| \,. \tag{6.4}$$

The accuracy of phone boundary placement for a given test set can be then approximated using the rate of phone boundary error which is below the chosen threshold which can be defined as

$$PBE_{ph,thr} = \frac{\sum_{i=1}^{N_{ph}} (PBE_{ph}[i] < thr)}{N_{ph}} \cdot 100 \tag{6.5}$$

where *ph* is phone/class identification, N_{ph} is the number of phone/class realizations, and *thr* is the value of chosen error threshold. Similarly, same procedure is applied for the computation of $PEE_{ph,thr}$ and $PLE_{ph,thr}$. Threshold values used for evaluations within this work were 5, 10, 20, or 30 ms respectively. All of these criteria can be computed with basic statistics for all phones, however, it is more common to do the evaluation over defined phone classes, which are generally language independent. The phone classes for English were defined according to [58], i.e. VOW - vowels, GLI - semivowels and glides, VFR - voiced fricatives, UFR - unvoiced fricatives, NAS - nasals, STP - stops, UST - unvoiced stops, and SIL - silence. Finally, *PronER* (*Pronunciation Error rate*) is also used to evaluate pronunciation detection accuracy

$$PronER = \frac{S}{N} \cdot 100 \tag{6.6}$$

where N is the total number of words in the reference set and S is the numbers of incorrectly recognized (substituted) pronunciation variants.

6.3 Baseline Phonetic Segmentation for English

The section describes baseline HMM-based phonetic segmentation and the task of finding the optimum acoustic model, the impact of extended pronunciation lexicon, and the accuracy of pronunciation variant detection when more variants are available in the lexicon. The experiments were performed using TIMIT corpus which is often used as a standard for the evaluation of phonetic segmentation for English. This also allowed us to compare obtained results with those published by other authors.

Experimental Setup

The experiments covering this task were realized with both predefined COMPLETE and CORE test sets available in TIMIT corpus. Only the phonetically-compact sentences (marked as SX sentences) and phonetically-diverse ones (marked as SI sentences) were the used. Following this selection criteria, the COMPLETE test set consisted of 50754 boundaries and the CORE set of 7215 boundaries. The summary of used data sets is presented in Table 3.10. TIMIT phone set was reduced from 61 to 48 phones for the purpose of creating the AM. This is a standard step that is the most often used for acoustic modelling. The set was then further reduced to 39 in order to perform the boundary scoring as this is a standard size of a phone set for English used in Kaldi recipes as well as by many other authors for their ASR systems [74]. We started with a standard approach to create the ASR (s5 recipe in Kaldi) and we optimized it with regards to improving the accuracy of automatic phonetic segmentation task. HMM topology consisted of 3 emitting states models for non-silence phones and 5 emitting states models for silence. Direct phone transcription, which includes also silence marks, was then used for the training the AM. Therefore, silence appearing in training graphs and silence boundaries were also scored, but optional silence were not.

6.3.1 Optimum AM for direct phonetic segmentation

The optimum choice of a proper AM had to be found. This task required using as precise of a transcription as possible, ideally using a a correct sequence of phones not words. As TIMIT contains transcriptions at a phone level, it allowed us to perform this step using this input for forced-alignment. When phonetic content is available, no phone needs to be recognized and PER is equal to 0 %. Obtained results are presented in the Table 6.1. Similarly to several other works (e.g. [82] or [134]), the highest accuracy were obtained for the simplest monophone AM, for both the CORE and COMPLETE test sets. Slightly lower accuracy of triphone- and DNN-based AMs might have been caused due to the fact that input features are taken from larger context, which yields to higher uncertainty when

	CORE SET				COMPLETE SET			
	5 ms	*10 ms*	*20 ms*	*30 ms*	*5 ms*	*10 ms*	*20 ms*	*30 ms*
mono	**29.16**	**52.79**	**83.08**	**93.00**	**29.00**	**52.71**	**82.79**	**92.63**
tri1	27.80	51.21	81.69	92.82	27.84	50.89	81.40	92.12
tri2	27.40	49.55	79.72	91.45	27.10	48.96	79.27	90.91
tri3	27.42	49.34	79.18	91.24	27.18	48.74	78.41	90.36
dnn	27.73	48.87	78.84	90.77	27.11	48.49	78.32	90.09

Table 6.1: Results of direct phonetic segmentation, $PER = 0$, $PCorr = 100$

	CORE SET				COMPLETE SET			
	5 ms	*10 ms*	*20 ms*	*30 ms*	*5 ms*	*10 ms*	*20 ms*	*30 ms*
mono144	31.05	54.57	82.51	92.17	31.37	54.67	81.90	91.73
mono288	**31.68**	**55.80**	84.70	**93.79**	**32.02**	**56.39**	**84.55**	**93.11**
mono432	30.45	54.73	**84.74**	93.74	31.03	55.32	84.46	93.06
mono720	29.76	53.50	83.53	93.35	29.95	53.70	83.48	92.99
mono1008	29.16	52.79	83.08	93.00	29.00	52.71	82.79	92.63
mono1440	28.18	51.50	81.80	92.82	28.13	51.31	81.80	92.30

Table 6.2: Optimization of monophone AM for direct phonetic segmentation ($PER = 0$, $PCorr = 100$)

determinaing a boundary position. Speaker dependent AMs achieved lower accuracy,

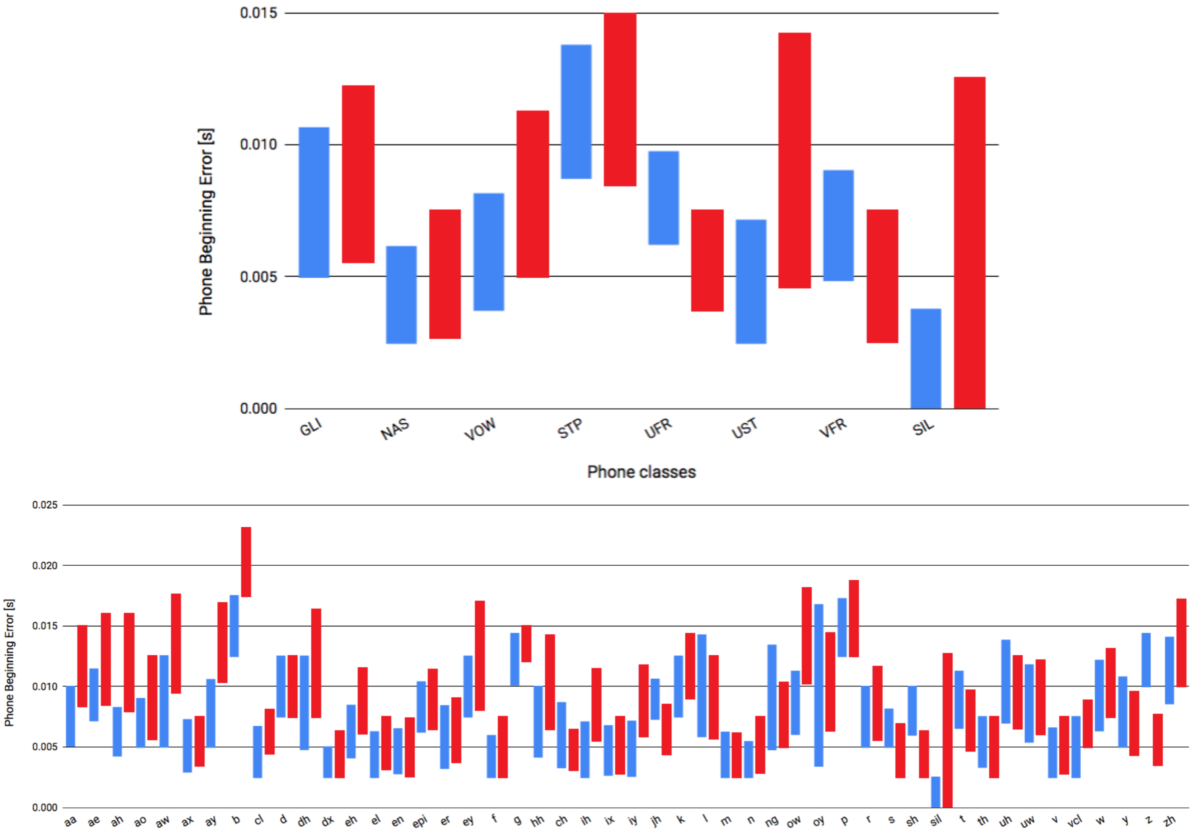


Figure 6.1: Phone Beginning Error PBE for particular phone lasses: blue - monophone system, red - DNN-based system

most likely, due to the limited amount of data available per speaker in the TIMIT corpus. Concerning the monophone AM, we looked for its optimized setup. Same as in other published works [82], it was confirmed that a smaller amount of Gaussian mixtures per state gives better results. The best results were achieved for 2 mixtures per state, see Table 6.2. The numbers in acronyms mono144, mono288, etc. in Table 6.1-6.2 represents the number of Gaussian components in whole HMM, e.g. 288 means 288 components with 2 mixtures per state, 3 emiting states per each monophone, and 48 phones in given HMM (2x3x48).

Finally, the distribution of values of *PBE* for particular phone classes is presented in Fig. 6.5. Particular bars describe distribution of *PBE* determined by 0.25 and 0.75 percentiles. Significantly worse results are observed for DNN system. However, the significant portion of the observed error can be attributed to the silence phone whereas the error increase for other phone classes is not so critical.

6.3.2 Phonetic segmentation with pronunciation variability

The second analysis describes the phonetic segmentation when exact phone sequence is not available and phonetic content is obtained from a pronunciation lexicon. This sitauation represents the most frequent scenario. However, the main issue is how well is the pronunciation variability covered in the lexicon and how the proper choice of word pronunciation variant influences the accuracy of the phonetic segmentation. We performed the experiments with 3 pronunciation lexicons: the first lexicon contained just *canonic pronunciations*, the second one contained all *pronunciation variants* observed in the TIMIT corpus, and the third one was based on merging previous two lexicons.

The lexicon containing all pronunciations which had appeared within phonetic transcription of TIMIT corpus (called further as *timit-variants*) was obtained from available transcriptions at the word and phone level. A significant majority of words from TIMIT had more than one pronunciation, so we could analyze also the ability of used AM to recognize the correct pronunciation variant for particular word realizations. In total, we obtained 19184 pronunciations for 6256 words, moreover, in some cases the number of pronunciation variants was very high (22 words had more than 20 pronunciations), as it is shown in more details in Table 6.6. This lexicon should also simulate using TIMIT corpus in a more realistic situation of informal speech when each word can have more pronunciations due to pronunciation variability in informal speaking style.

Obtained results are shown in Tabs 6.3-6.5 and a significant decrease of *PER* was observed when lexicon contained pronunciation variants. Further, the usage of more advanced AM (DNN-based one) contributed to further decrease of achieved *PER* below 10%. Consequently, it means the increase of *PCorr*, i.e. more than 92% of all phones were

		PER	*PCorr*	*5 ms*	*10 ms*	*20 ms*	*30 ms*
CORE	mono	32.58	71.43	23.94	43.54	72.39	85.82
	tri1	32.8	71.58	23.15	42.23	70.28	84.19
	tri2	32.55	71.64	22.96	40.82	67.83	83.45
	tri3	32.46	71.57	23.92	40.41	66.08	81.70
	dnn	31.88	71.45	23.67	40.14	65.28	80.60
	mono288-dnn	31.88	71.45	25.78	45.37	72.01	84.54
COMPLETE	mono	31.15	72.28	23.92	43.23	72.34	85.78
	tri1	31.79	72.18	23.23	41.96	70.18	84.21
	tri2	31.45	72.22	23.00	40.50	67.82	83.16
	tri3	31.3	72.24	23.47	40.41	66.33	81.20
	dnn	30.52	72.28	23.43	40.32	65.83	80.59
	mono288-dnn	30.52	72.28	26.39	46.38	72.71	84.93

Table 6.3: Phonetic segmentation with canonic lexicon

		PER	*PCorr*	*5 ms*	*10 ms*	*20 ms*	*30 ms*
CORE	mono	12.24	89.69	28.77	51.61	82.26	92.58
	tri1	11.49	90.85	27.19	50.18	80.70	92.01
	tri2	11.16	91.14	26.93	48.59	78.89	90.71
	tri3	10.24	91.91	27.51	48.45	78.03	90.59
	dnn	**9.58**	92.03	27.64	48.55	78.03	90.05
	mono288-dnn	**9.58**	92.03	**31.28**	**54.94**	**83.81**	**93.09**
COMPLETE	mono	12.06	89.62	28.82	52.11	82.25	92.30
	tri1	11.89	89.83	27.58	50.28	80.81	91.65
	tri2	11.17	91.19	26.88	48.41	78.50	90.39
	tri3	10.75	91.46	27.04	48.10	77.68	89.90
	dnn	**10.00**	92.06	27.16	48.28	77.73	89.55
	mono288-dnn	**10.00**	92.06	**31.91**	**55.98**	**84.17**	**92.93**

Table 6.4: Phonetic segmentation with TIMIT-variant lexicon

		PER	*PCorr*	*5 ms*	*10 ms*	*20 ms*	*30 ms*
CORE	mono	12.43	89.48	28.79	51.69	82.25	92.64
	tri1	11.74	90.64	27.25	50.17	80.72	92.08
	tri2	11.31	91.02	26.97	48.64	78.91	90.76
	tri3	10.42	91.75	27.48	48.46	78.08	90.56
	dnn	**9.76**	91.88	27.65	48.51	77.99	90.00
	mono288-dnn	**9.76**	91.88	**31.33**	**55.00**	**83.84**	**93.12**
COMPLETE	mono	12.40	89.28	28.83	52.08	82.25	92.31
	tri1	11.45	90.92	27.60	50.29	80.83	91.69
	tri2	11	91.2	26.90	48.44	78.51	90.42
	tri3	10.23	91.84	27.04	48.11	77.64	89.90
	dnn	**9.28**	92.17	27.12	48.22	77.63	89.44
	mono288-dnn	**9.28**	92.17	**31.92**	**55.97**	**84.14**	**92.93**

Table 6.5: Phonetic segmentation with canonic lexicon extended by TIMIT variants

No. of pronunciation variants	1	2	3-5	6-10	11-20	< 20
No. of words	631	3372	1516	637	78	22

Table 6.6: Lexicon timit-variants - statistics

correctly identified, however, the accuracy of boundary determination slightly decreased when DNN-based system was used. On the other hand, when the recognized phone sequence is realigned with optimized monophone system with 288 Gaussian components (acronym mono288-dnn), both the best *PER* and boundary placement accuracy were achieved [131].

6.3.3 Pronunciation recognition

When pronunciation lexicons contain a very high number of pronunciation variants, correct detection of the proper pronunciation variant is a very important task and phonetic segmentation in this scenario can also serve the purpose of detecting proper pronunciation variants within an analyzed utterance. It can then play an important role in the research focused on pronunciation variability so the correctness of pronunciation variant selection was analyzed at the end as well.

In fact, the choice of correct pronunciation was already quantified to some degree by looking at the the decrease in *PER* described in previous section. However, for many words we had a rather high amount of pronunciation variants which could be very important feature of such a system. From the results described in Table 6.3.3, we can observe a significant decrease in *PronER* when more advanced acoustic modelling are used. The same held true for lexicons. The best results were obtained with DNN-based system, where we observed a significant decrease in *PronER*; 76.34% were obtained for basic monophone system and CORE test set, while 31.89% weas achieved for DNN-based system. The contribution of GMM-HMM systems with triphone-based models was seen too. The same trend in obtained results was also observed for the COMPLETE set.

6.3.4 Summary

The implementation of the HMM-based phonetic segmentation was presented together with the analysis of various acoustic modelling techniques on the final accuracy of phone-boundaries determination. The evaluations were performed with TIMIT database and they proved the contribution of advanced acoustic modelling when the task was to choose the proper pronunciation variant. We achieved more than 92% correctness for phone recognition within forced-alignment with the DNN-HMM system. The improvement of phone boundary placement was also observed in the second step with an optimized mono-

		canonic		timit		canonic+variants	
		PER	PronER	PER	PronER	PER	PronER
CORE	mono	32.58	76.34	12.24	39.48	12.43	40.18
	tri1	32.80	76.28	11.49	37.82	11.74	38.46
	tri2	32.55	76.28	11.16	35.97	11.31	36.54
	tri3	32.46	76.28	10.24	33.48	10.42	34.06
	dnn	31.88	**76.34**	9.58	**31.44**	9.76	**31.89**
COMPLETE	mono	31.15	74.22	12.06	40.39	12.40	41.44
	tri1	31.79	74.21	11.89	37.06	11.45	37.87
	tri2	31.45	74.22	11.17	35.77	11.00	36.60
	tri3	31.30	74.21	10.75	33.82	10.23	34.56
	dnn	30.52	**74.22**	10.00	**31.46**	9.28	**32.19**

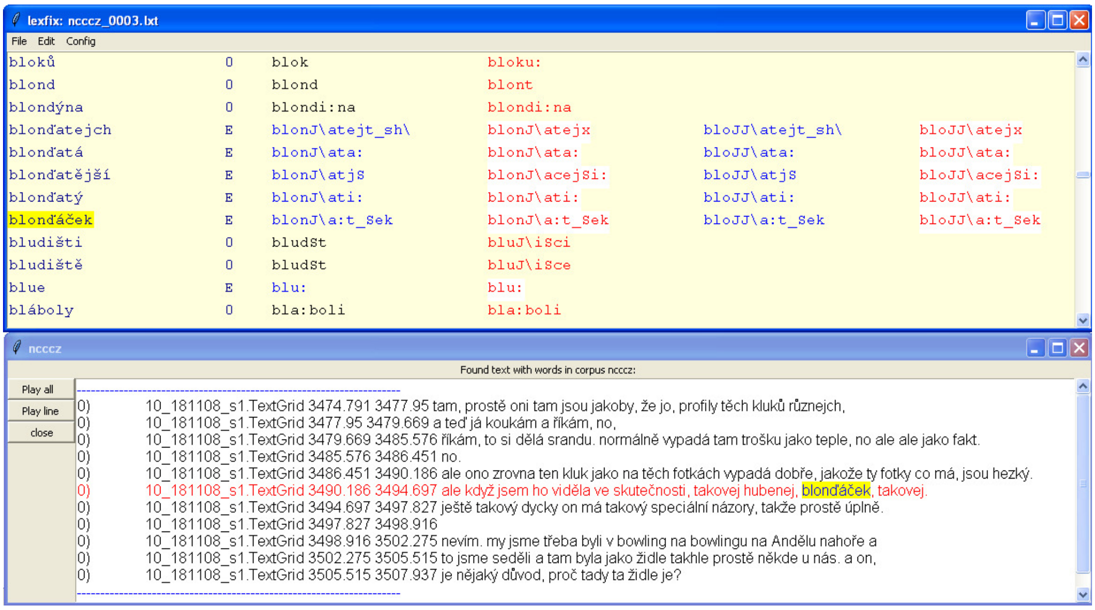
Table 6.7: Pronunciation variant recognition

phone GMM-based system; 83.84% of phone beginning boundaries were determined with the threshold smaller than 20 ms, and the error reached 93.12% for the threshold smaller than 30 ms. These results were obtained without any further boundary correction, as that was not the goal in our applications, and the referenced authors did not employ such techniques either.

6.4 Phonetic segmentation of Czech casual speech

The accuracy of the HMM-based forced-alignment technique used for phonetic segmentation relies on the quality of acoustic data. It also depends strongly on the accuracy of input phonetic contents. This section analyses the accuracy of phonetic segmentation performed on NCCCz which contains informal speech of strong spontaneous nature. It influences the character of the produced speech at various levels, mainly at the level of rather free pronunciation.

Phonetic content of utterances is transcribed usually at orthographical level and can be obtained by grapheme-to-phoneme conversion or from a pronunciation lexicon. The basic lexicon with canonical pronunciations of words in NCCCz had to be created. The description of this procedure and a tool supporting pronunciation check of lexicon items is described below. To conduct the experiment of phonetic segmentation on NCCCz corpus, the lexicon must cover the pronunciation variability, which was achieved by including additional pronunciation variants. The lexicon with various pronunciation variants which covers speech variability caused by processes of co-articulation, assimilation or reduction had to be used when phone boundaries were to be determined for spontaneous and informal speech, higher diversity of language dialects, as well as in other situations when the level of pronunciation variability is rather high. Such a lexicon can be obtained manually (for some very specific situations) or automatically (to extend regular pronunciations by



```
bloků          0   blok               bloku:
blond          0   blond              blont
blondýna       0   blondi:na          blondi:na
blonďatejch    E   blonJ\atejt_sh\    blonJ\atejx       bloJJ\atejt_sh\   bloJJ\atejx
blonďatá       E   blonJ\ata:         blonJ\ata:        bloJJ\ata:        bloJJ\ata:
blonďatější    E   blonJ\atjS         blonJ\acejSi:     bloJJ\atjS        bloJJ\acejSi:
blonďatý       E   blonJ\ati:         blonJ\ati:        bloJJ\ati:        bloJJ\ati:
blonďáček      E   blonJ\a:t_Sek      blonJ\a:t_Sek     bloJJ\a:t_Sek     bloJJ\a:t_Sek
bludišti       0   bludSt             bluJ\iSci
bludiště       0   bludSt             bluJ\iSce
blue           E   blu:               blu:
bláboly        0   bla:boli           bla:boli
```

```
------------------------------------------------------------------------
0)      10_181108_s1.TextGrid 3474.791 3477.95 tam, prostě oni tam jsou jakoby, že jo, profily těch kluků různejch,
0)      10_181108_s1.TextGrid 3477.95 3479.669 a teď já koukám a říkám, no,
0)      10_181108_s1.TextGrid 3479.669 3485.576 říkám, to si dělá srandu. normálně vypadá tam trošku jako teple, no ale ale jako fakt.
0)      10_181108_s1.TextGrid 3485.576 3486.451 no.
0)      10_181108_s1.TextGrid 3486.451 3490.186 ale ono zrovna ten kluk jako na těch fotkách vypadá dobře, jakože ty fotky co má, jsou hezký.
0)      10_181108_s1.TextGrid 3490.186 3494.697 ale když jsem ho viděla ve skutečnosti, takovej hubenej, blonďáček, takovej.
0)      10_181108_s1.TextGrid 3494.697 3497.827 ještě takový dycky on má takový speciální názory, takže prostě úplně.
0)      10_181108_s1.TextGrid 3497.827 3498.916
0)      10_181108_s1.TextGrid 3498.916 3502.275 nevím. my jsme třeba byli v bowling na bowlingu na Andělu nahoře a
0)      10_181108_s1.TextGrid 3502.275 3505.515 to jsme seděli a tam byla jako židle takhle prostě někde u nás. a on,
0)      10_181108_s1.TextGrid 3505.515 3507.937 je nějaký důvod, proč tady ta židle je?
------------------------------------------------------------------------
```

Figure 6.2: Illustrative example of the work with the LexFix tool

particular phone substitutions or reductions on the basis of defined rules.

6.4.1 The NCCCz lexicon

The Nijmegen Corpus of Casual Czech was created to provide a corpus of Czech containing high-quality recordings from naturally occurring interactions which would be suitable for a detailed analysis of spontaneous speech. For the purpose of further studies and developments, the orthographic transcription of records had to be created using a proper pronunciation lexicon. Generally, a lexicon always represents an important component which has a significant impact on the accuracy of the target ASR system. It is especially important in the case of spontaneous or casual speech recognition for which the process of coarticulation, assimilation and reduction often appears. Due to the very informal speaking style resulting in the appearance of many rare or non-standard words, available lexicon containing regular canonical pronunciation had to be corrected manually before being applied to NCCCz. The first version was created using grapheme-to-phoneme conversion rules that have been verified for another lexicon for Czech [105].

The automatically generated pronunciation contained a large amount of incorrect pronunciations mainly for foreign words as well as for above-mentioned non-standard ones. The correction (editing) of the pronunciation lexicon was clearly needed. For this purpose, the *LexFix* tool for lexicon editing was modified to allow working with the orthographic transcription and listening to a recorded utterance at the same time. In general, the tool was created to correct pronunciation of any word form, not only the ones appearing in casual or informal speech. The ability to search for a word together with the neighbouring context in a huge corpus was necessary. It might also be very difficult to decide about the pronunciation of a non-standard word without listening to it, so it was necessary to play the particular sentence the word appeared in. The illustrative example of the work with the *LexFix* tool is at Fig. 6.2.

During these checks, reduced pronunciations (e.g. "nějaký" vs. "ňáký") were not marked so the current lexicon contains only canonical pronunciation with a small amount of pronunciation variants. Finally, the created pronunciation lexicon for NCCCz contains approx 30 000 word forms. To capture the variability of informal spontaneous speech pronunciation, created canonical lexicon was completed by adding more pronunciation variants in next steps. For this purpose, the rules described in Sec. 3.4.3 were used and brief summary is presented in Table 6.8.

		No. of pronunciation variants					
lexicon type	total size	1	2	3-5	6-10	11-20	< 20
lexicon_A	29077	27812	625	5	-	-	-
lexicon_B	136065	7311	6431	7641	4340	2352	107974

Table 6.8: NCCCz Lexicons variants - statistics

The analysis of phonetic segmentation accuracy was realized using the following two variants of pronunciation lexicons:

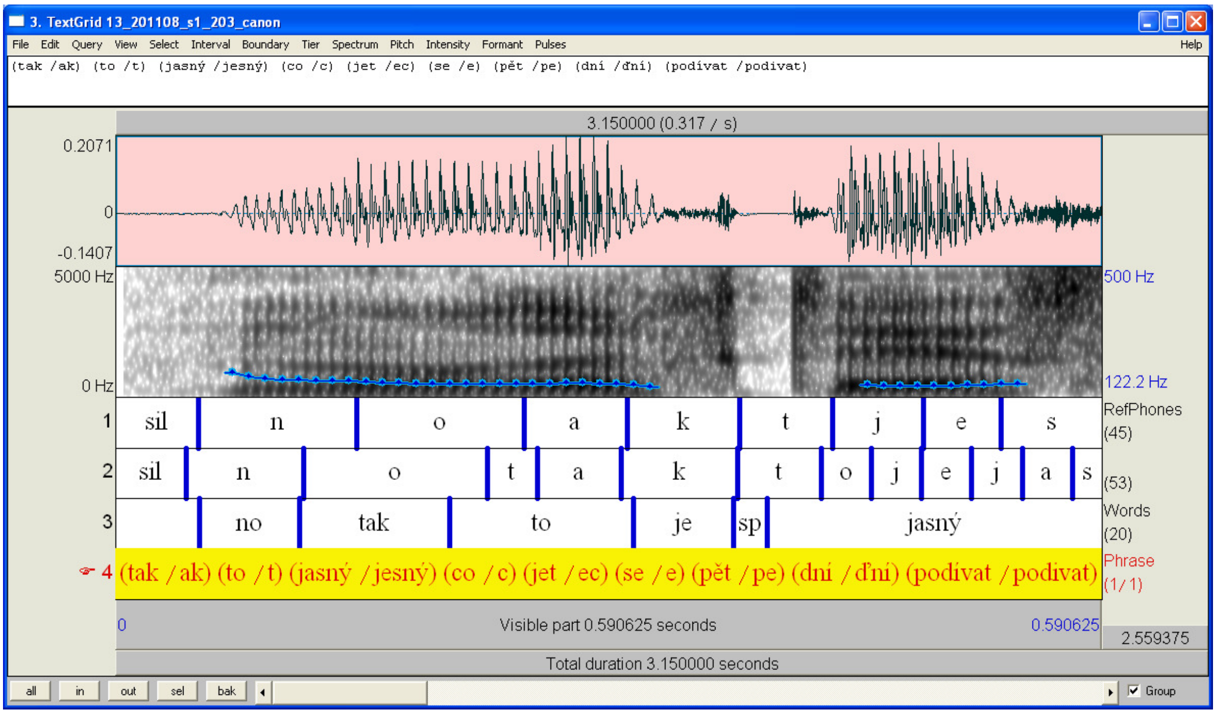
- Lexicon_A - the lexicon with *canonical pronunciation*,
- Lexicon_B - the lexicon with *canonical pronunciation* extended with rule based pronunciation variants.

Experimental Setup

The acoustic models used for experiments described in this section were trained on NCCCz train data set, see Table 3.6. The experiments were realized with the following AMs: *mono*, *tri1*, *tri2*, *tri3*, *dnn*. One difference, however, was the model denoted as *dnn_mono*, which was DNN system (2 hidden layers and 1024 hidden units) trained on monophones as the targets, and not the senons, as is usually the case. To demonstrate the quality of these acoustic models, Tabs. 3.11 and 3.12 present the WERs for a LVCSR task. Phonetic segmentation experiments were carried out using utterances from the NCCCz which were summarized in more detail in Table 3.7.

6.4.2 Optimum AM for direct phonetic segmentation

A manually created phonetic transcription was used as an input for HMM-based forced alignment. The direct phonetic transcription represents an ideal case when proper pronunciation is selected from pronunciation variants in lexicon. The demonstration of the

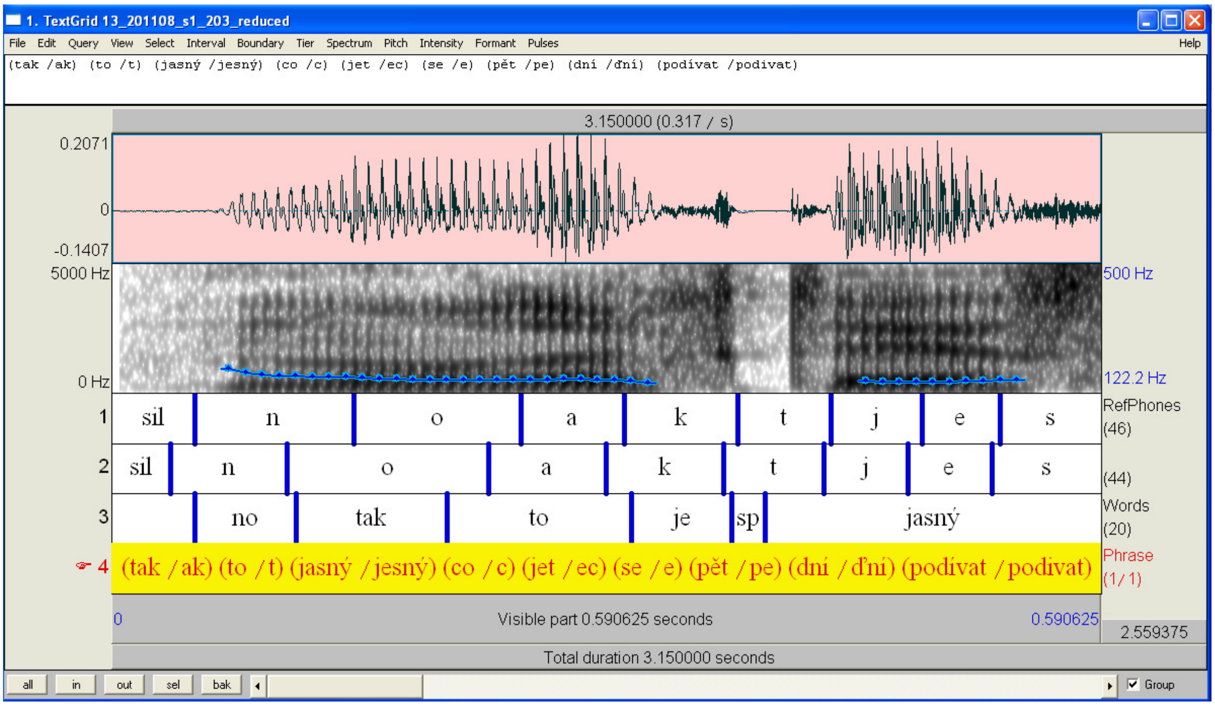


3. TextGrid 13_201108_s1_203_canon

(tak /ak) (to /t) (jasný /jesný) (co /c) (jet /ec) (se /e) (pět /pe) (dní /ďní) (podívat /podivat)

File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
3.150000 (0.317 / s)
0.2071
0
-0.1407
5000 Hz
0 Hz
500 Hz
122.2 Hz
1 sil n o a k t j e s
RefPhones (45)
2 sil n o t a k t o j e j a s
(53)
3 no tak to je sp jasný
Words (20)
4 (tak /ak) (to /t) (jasný /jesný) (co /c) (jet /ec) (se /e) (pět /pe) (dní /ďní) (podívat /podivat)
Phrase (1/1)
0
Visible part 0.590625 seconds
0.590625
2.559375
Total duration 3.150000 seconds
all in out sel bak
Group

Figure 6.3: Illustrative example of phonetic segmentation results with the canonical pronunciation.



1. TextGrid 13_201108_s1_203_reduced

(tak /ak) (to /t) (jasný /jesný) (co /c) (jet /ec) (se /e) (pět /pe) (dní /ďní) (podívat /podivat)

File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
3.150000 (0.317 / s)
0.2071
0
-0.1407
5000 Hz
0 Hz
500 Hz
122.2 Hz
1 sil n o a k t j e s
RefPhones (46)
2 sil n o a k t j e s
(44)
3 no tak to je sp jasný
Words (20)
4 (tak /ak) (to /t) (jasný /jesný) (co /c) (jet /ec) (se /e) (pět /pe) (dní /ďní) (podívat /podivat)
Phrase (1/1)
0
Visible part 0.590625 seconds
0.590625
2.559375
Total duration 3.150000 seconds
all in out sel bak
Group

Figure 6.4: Illustrative example of phonetic segmentation results with the reduced.

AM type	*PER*	*PCorr*	*5 ms*	*10 ms*	*20 ms*	*30 ms*
mono_135	0	98.17	34.40	57.74	79.70	87.61
mono_270	0	100	39.86	64.28	85.13	92.21
mono_405	0	100	37.67	61.69	83.33	91.69
mono_540	0	100	37.89	61.15	83.15	91.32
mono_675	0	100	36.58	60.68	83.25	91.22
mono_810	0	100	35.45	58.98	82.44	91.17
mono_945	0	100	34.11	58.12	81.94	90.68
mono_1080	0	100	36.21	59.37	82.96	91.14
mono_1215	0	100	35.20	58.76	82.34	90.95
mono_1350	0	100	34.36	58.12	82.46	91.12
tri1	0	100	37.25	60.88	83.60	91.93
tri2	0	98.17	34.17	57.71	82.04	90.93
tri3	0	100	33.10	56.14	80.59	89.39
dnn	0	100	3229	5459	7950	8873
dnn_mono	0	100	36.21	60.66	84.58	92.33

Table 6.9: Results of direct phonetic segmentation of casual Czech

impact of proper pronunciation selection on phonetic segmentation accuracy for casual speech is illustrated in Figs. 6.3 and Fig. 6.4. The figures demonstrate the importance of proper pronunciation on phonetic segmentation accuracy for words with strong pronunciation reduction.

Obtained results for analysed AMs are summarized in Table 6.9. In the case of casual speech, the best results were obtained with *dnn_mono* AM at the level of 30 ms threshold and the optimized *mono_270* system achieved the best results for 5, 10, 20 ms thresholds. Similar results were observed for English, where the best results were achieved by using *mono* system. The numbers used in acronyms in Table 6.9 (mono135, mono270, mono405 etc.) represents the total number of Gaussian components in the model. The speaker independent DNN system, which was trained on the same features as the *mono* system and monophones as targets, outperformed the speaker-dependent *dnn* system which was trained on stacked fMLLR features.

6.4.3 Segmentation with reduced pronunciation

The next step was to compare the accuracy of phonetic segmentation when lexicon with canonical pronunciation is used as well as when the pronunciation variability is captured. The results for phonetic segmentation with canonical pronunciation are in Table 6.10 and for the segmentation with lexicon with additional variants are in Table 6.11. The *dnn_mono* system achieved the best segmentation accuracy of approx. 90% with the 20/30ms thresholds for both lexicons. The best segmentation accuracy of approx. 37% with the 5ms threshold was archived by *mono_270*. Additional *PER* reduction was achieved, when the phone sequence was recognized with *dnn* model and then realigned with *dnn_mono* or *mono* models. Although the presented results were obtained in experiments performed on a small evaluation subset of manually segmented utterances, the results had demonstrated the contribution of including information about pronunciation reduction into the lexicon.

AM type	PER	PCorr	5 ms	10 ms	20 ms	30 ms
mono_270	16.51	93.05	37.19	60.27	81.53	88.65
mono_405	16.36	93.07	35.42	58.63	80.38	89.08
tri1	16.65	93.12	35.19	57.21	80.81	89.40
tri2	16.56	93.12	32.42	54.52	78.87	87.77
tri3	16.44	93.17	31.46	53.57	77.62	86.89
dnn	16.26	93.20	31.29	53.79	77.66	86.50
dnn_mono	16.04	93.17	34.81	58.31	82.07	89.85

Table 6.10: lexicon A - canonical pronunciation

AM type	PER	PCorr	5 ms	10 ms	20 ms	30 ms
mono_270	18.87	88.42	37.37	60.76	82.02	89.11
mono_405	18.48	89.03	34.84	58.31	80.80	89.10
tri1	16.87	91.73	35.65	57.84	81.45	90.03
tri2	16.85	91.59	32.93	55.14	79.69	88.37
tri3	16.14	92.20	32.00	54.24	78.31	87.46
dnn	15.83	92.44	31.21	53.97	78.37	86.52
dnn_mono	17.36	89.93	34.92	58.60	82.81	90.43
dnn-mono_270	15.83	92.44	37.38	60.67	81.96	89.00
dnn-dnn_mono	15.83	92.44	34.95	58.67	82.62	90.32

Table 6.11: lexicon B - canonical pronunciation extended with rule based pronunciation variants

6.4.4 Reduced pronunciation recognition

The final step was to analyze the accuracy of pronunciation variant recognition on casual Czech speech. In the case of casual speech, it is more difficult because the acoustic realization is often strongly irregular and influenced by an informal speaking style. Moreover, when the lexicon covering possible pronunciation variability is obtained using many predefined rules applied to all words, it contains consequently a rather big amount of pronunciation variants where many of them could be acoustically very similar as well as the probability of appearance for many of them could be rather low. By looking at the obtained results presented in Table 6.12, a significant improvement in accuracy on the level of *PronER* was achieved with a DNN based system. The system outperformed the GMM-HMM *mono*, *tri1*, *tri3* systems and also speaker independent *dnn_mono*.

system	lexicon_A	lexicon_B
mono_270	37.28	44.98
mono_405	37.28	44.13
tri1	37.09	38.59
tri2	37.09	38.69
tri3	37.09	37.37
dnn	37.09	37.56
dnn_mono	37.28	42.44

Table 6.12: Pronunciation variant recognition for casual Czech

6.5 Impact of AF for phonetic segmentation

The presented studies have analyzed the contribution of various types of AMs and lexicons where the main goal was to maximize accuracy of the automatic phonetic segmentation. It means that matched training and testing data were used (i.e. training and testing data were both form NCCCz). In this section, a more realistic scenario is analyzed, when the AM used for HMM-based phonetic segmentation was trained on data from the SPEECON corpus. It means that NCCCz acoustic conditions were not seen during the AM training, which mean that the automatic phonetic segmentation was tested on purely unseen data. Such mismatch in speaking style represented an additional challenge for the HMM-based phonetic segmentation task. To minimize the both the acoustic and speaking style mismatches, the idea was to analyze the potential contribution of AF-based TANDEM system for this task. The developed system described in the previous section was used now for comparison purposes.

Experimental Setup

Finally, three types of AMs were used in the experiments. The first AM was based on GMM-HMM architecture and the system was described in section 3.4.2. The second one was based on AF-TANDEM architecture and the system was described in section 5.2.2. Both GMM-HMM and AF-Based Tandem systems were trained only on speech data from SPEECON database. The *mono* and *tri1* types of AM were used for the analyses. The third AM was trained on matched acoustic conditions (i.e. on NCCCz train set) and the training procedure for this AM was described in section 3.4.2. The NCCCz test set described in the Table 3.6 was used for the evaluation and both *Lexicon A* and *Lexicon B* (NCCCz lexicons) were considered during these experiments.

Results & Discussion

The achieved results are summarized in Tab 6.13. The *tri1* GMM-HMM system achieved the best segmentation accuracy of approx. 87% with the 30ms threshold. This was bet-

AM train set	test set	AM type	PER	5ms	10ms	20ms	30ms
Lexicon_A							
speecon	ncccz_test	mono	15.94	26.37	46.23	75.72	86.67
speecon	ncccz_test	tri1	15.89	26.92	48.45	78.13	87.14
speecon	ncccz_test	af_mono	15.48	26.13	46.88	73.26	83.65
speecon	ncccz_test	af_tri	16.03	24.93	44.29	73.11	83.86
ncccz	ncccz_test	mono	16.09	33.63	56.67	80.10	88.42
ncccz	ncccz_test	tri1	16.56	34.22	56.51	80.85	89.35
Lexicon_B							
speecon	ncccz_test	mono	20.19	26.26	45.52	74.92	86.02
speecon	ncccz_test	tri1	18.49	26.87	48.03	77.37	86.82
speecon	ncccz_test	af_mono	19.82	25.27	45.49	72.67	83.44
speecon	ncccz_test	af_tri	18.98	25.34	44.93	73.44	84.77
ncccz	ncccz_test	mono	18.07	33.76	56.81	80.32	88.57
ncccz	ncccz_test	tri1	17.07	34.80	57.44	81.69	89.74

Table 6.13: The results for phonetic segmentation in mismatched scenario

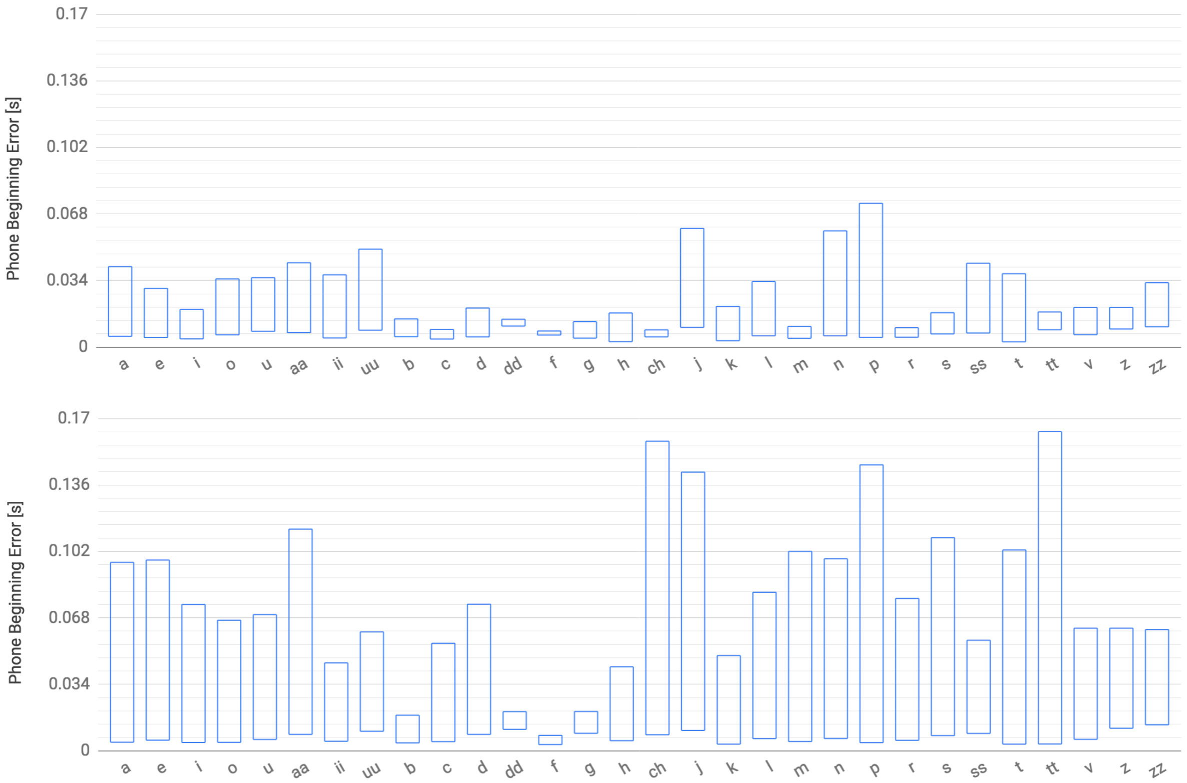


Figure 6.5: Phone Beginning Error *PBE*. The first figure: GMM-HMM monophone system. The second figure: AF-based Tandem monophone system. Lexicon A setup.

ter than both the *mono* GMM-HMM and *mono/tri1* AF-based TANDEM systems. The *tri1* GMM-HMM system achieved around 2% worse results with 30*ms* threshold when compared to the results achieved with the *tri1* model trained on NCCCz. The significant degradation was seen for both GMM-HMM and AF-Based TANDEM systems for more strict thresholds of $5ms$, $10ms$ and $20ms$. The positive impact of more pronunciation variants *Lexicon B* on the accuracy of phonetic segmentation was observed for the matched AM. In the case of GMM-HMM and AF-Based TANDEM systems, using the more generic lexicon showed a negative impact. The comparison between GMM-HMM and AF-Based TANDEM system on the level of Phone Beginning Error criteria is presented in the Fig. 6.5. The worse performance of AF-Based TANDEM system could be explained by the worse generalization of AF to different speaking styles encountered in NCCCz. The degradation of AF estimation of approx. 25% for average *FAcc* was measured. The authors in [126] reached the same conclusion.

6.5.1 Summary

In this section, the pilot analysis of HMM-based phonetic segmentation accuracy with regards to the usage of canonical and reduced pronunciations were performed. The experiments were done with the speech from Nijmegen Corpus of Casual Czech (NCCCz), which contains speech of a very strong and spontaneous nature. The results demonstrated the significance of pronunciation reduction on the proper acoustic modelling of spontaneous speech (applied currently on phonetic segmentation).

The proper lexicons were created to ameliorate this issue. The LexFix tool, which was created as part of this thesis, represents another the important contribution of this work because it supports general lexicon editing with the possibility to locate the word and its neighbouring context in a very large corpus, and listening to it as the same time. The reduction of the pronunciation is supposed to be solved automatically at further steps of this wider research. On the other hand, the procedure of looking for the reduction rules is also supposed to be supported by the NCCCz data together with the possible listening of particular occurrences of lexicon items in the corpus.

The two stage forced-alignment consisting of combination of speaker-dependent DNN-HMM system and DNN-HMM trained on monophone targets was analyzed on casual speech NCCCz test set and showed to achieve results comparable to English read speech.

In the end, the performance of GMM-HMM and AF-Based TANDEM systems was analyzed in the mismatch conditions. The benefits of the AF-Based TANDEM systems could not be confirmed for automatic phonetic segmentation.

Chapter 7

Conclusions

The general goal of this thesis was to study basic the properties of articulatory features, their estimation techniques and their potential applications in ASR systems and for the task of phonetic segmentation. Particular conclusions were already discussed in more details within previous chapters but the most important contributionss of this thesis can be summarized as follows.

- The general properties of AF were studied and the state-of-the-art for their possible applications was presented.
- AF classes for Czech were defined, including the mapping of phones to particular AF categories. Only a minimal number of works described an up-to-date research on AF for Czech, moreover, majority of them deal with other target applications (e.g. speech synthesis). In addition, they often use slightly different approaches. A similar unification of AF classes was done also for several other East-European languages, namely for Slovak, Polish, Hungarian, and Russian.
- DNN-based estimation of AF was optimized. This task was done for Czech, English, Slovak, Polish, Hungarian, and Russian languages. The optimum temporal context as an input for a DNN was estimated to be between $210 \div 310$ ms for all languages. The modern techniques of feature extraction such as DCT-TRAP, stacked cepstral feature, MFCC-LDA-MLLT, or FMLLR features were analyzed for the classification of AF features. The fMMLR feature proved to have a positive impact on AF estimation. The Czech AF classes were estimated with the average *FAcc* of around 90%.
- The visualization of estimated AF in the Praat environment was prepared. It is expected to help in a study of AF estimation accuracy as well as in the research of phonetics and other fields where the articulation is analyzed.

- Incorporation of AF into the phone recognition, AF-TANDEM-based LVCSR, as well as phonetic-segmentation was implemented as the first steps to improve the recognition of spontaneous and informal speech. These steps were focused mainly on Czech and English languages.

- The first experiments were related to the above mentioned tasks on corpora with a formal read speech (database SPEECON, TIMIT, SpeechDat), and on the corpora with the speech containing a higher level of background noises (car speech data), and finally on spontaneous speech (data containing technical lectures) and casual speech. The design of Czech casual speech recognition system with the focus on the optimization of the acoustic and language models was presented. Concerning the obtained results, the best setup was achieved for the DNN-HMM system with merged language model and pronunciation variation modelling. It achieved 58.4% *WER*, which is comparable to the results presented by other authors. The lexicon with manually corrected canonical pronunciations improved the results by about 1%, in terms of the *WER*. The built system was also evaluated on other spontaneous data (lecture recordings, which were slightly more formal) where it achieved somewhat better *WER* of 37.2%. In addition, this system was evaluated on the of formal read speech recognition where the WER of 14.7% was achieved. The observed margin between the casual and formal speech recognition illustrated the challenge for the research in the field of more informal speech recognition.

- The review of AF contribution to phone recognition and ASR performances under various speaking style was presented. It was done using an AF-based Tandem ASR system. The positive impact of AF-Based TANDEM system was observed for standard (*mono*) and triphone (*tri1*) systems for both languages and confirm the previous results achieved by other authors. The next improvement was achieved using the ASR combination of GMM-HMM or DNN-HMM and AF-Based TANDEM systems. The most important result was the 17% *PER*, which was achieved for a combined DNN-HMM system that was trained with and without AF features. The AF-based Czech TANDEM surpassed the GMM-HMM (*mono*) and triphone (*tri1*) systems for the task of causal speech recognition.

- The impact of various types of AM on automatic phonetic segmentation was analyzed for two speaking styles on TIMIT for English read speech, and NCCCz with Czech casual speech. The two stage forced-alignment consisting of a combination of DNN-HMM and optimized monophone GMM-HMM-based or DNN-HMM based system was proposed in this thesis. The positive impact on the level of phone boundary determination was observed for both read English speech and casual Czech speech test sets. The best phone boundaries accuracy on the TIMIT was around

93% for the 30ms criteria and a 90% accuracy was achieved on NCCCz test set. The positive contribution of using a lexicon with reduced pronunciation variants was confirmed. The AF-Based TANDEM system was analyzed in a mismatched setup but the system did not improve the results.

- Concerning the implementation as as a by-product of this thesis, Czech ASR LVCSR system was implemented using the modern Kaldi toolkit and the recipes for all available Czech corpora in our lab such as CZKCC, NCCCz, SPEECON, SpeechDat, CzLecDSP, CtuTest were created. The CtuCopy feature extraction tool was extended by cepstral normalization techniques.
- Concerning the experimental part, the processing of the NCCCz corpus was finalized. The data sets of NCCCz were cleaned-up and converted to a format suitable for being used in Kaldi recipes and they are now available at *http://www.mirjam ernestus.nl/Ernestus/NCCCz/index.php.*
- A reviewed canonical lexicon was added to the NCCCz corpus. The corpus now contains manually checked pronunciations of 29077 words. The LexFix tool was adapted for this purpose and it can be used for other similar tasks. Finally, the most important pronunciation variability was included and 630 words have additional pronunciation variants.

Concerning the final summary, the thesis presented the study of estimation AF features with the focus on the accuracy and possible contribution for speech processing of Czech casual speech. The realized analyses for Czech language confirmed that AF feature contained complementary information to standard cepstral features and consequently they can contribute to the improvement of recognition accuracy with TANDEM-based system using monophone and triphone GMM-HMM models. Similar results were presented also by other authors for English language. Unfortunately, for more advanced GMM-HMM systems based on stacked cepstral features followed by LDA/MLLT, or systems based on a DNN-HMM architecture, the complementary information in the form of AF features was not found to be of used. It is our hypotheses that this happened because these systems use features which include longer temporal context, or is is the case for DNN-HMM systems, the deep model extracts information similar to the one carried by AF features.

On the other hand, the proposed estimation of AF features seems to be precise, so further work on applying AF for speech processing task is a possibility. Further research could be focused on improving current speech application such as computer assisted pronunciation training or text-to-speech conversion systems. The usage of AF features for biomedical clinical application, such as for the analysis of disordered speech, represents a potentially interesting application area.

Appendix A

Summary of articulatory features mapping for all languages

The following tables provide the summary of articulatory feature per particular Slovak, Polish, Hungarian, Russia phones in X-SAMPA phonetic alphabet.

Phones	*Place_con*	*Manner_con*	*palatalization*	*Sonor*	*Voicing*	*Manner_vow*	*Place_vow*
C	palatal	fricatives	nil	-	-	nil	nil
F	labiodental	nasals	nil	+	+	nil	nil
J	palatal	nasals	nil	+	+	nil	nil
J:	palatal	nasals	nil	+	+	nil	nil
J	palatal	stop	nil	-	+	nil	nil
J	palatal	stop	nil	-	+	nil	nil
L	palatal	lateral	nil	+	+	nil	nil
N	velar	nasals	nil	+	+	nil	nil
S	postalveolar	fricatives	nil	-	-	nil	nil
S’:	postalveolar	fricatives	palatalized	-	-	nil	nil
S:	postalveolar	fricatives	nil	-	-	nil	nil
Z	postalveolar	fricatives	nil	-	+	nil	nil
Z:	postalveolar	fricatives	nil	-	+	nil	nil
b	bilabial	stop	nil	-	+	nil	nil
b’	bilabial	stop	palatalized	-	+	nil	nil
b:	bilabial	stop	nil	-	+	nil	nil
c	palatal	stop	nil	-	-	nil	nil
c:	palatal	stop	nil	-	-	nil	nil
d	prealveolar	stop	nil	-	+	nil	nil
d:	prealveolar	stop	nil	-	+	nil	nil
dZ	postalveolar	affricates	nil	-	+	nil	nil
dz	prealveolar	affricates	nil	-	+	nil	nil
dz:	prealveolar	affricates	nil	-	+	nil	nil
f	labiodental	fricatives	nil	-	-	nil	nil

Phones	*Place_con*	*Manner_con*	*palatalization*	*Sonor*	*Voicing*	*Manner_vow*	*Place_vow*
f'	labiodental	fricatives	palatalized	-	-	nil	nil
f:	labiodental	fricatives	nil	-	-	nil	nil
g	velar	stop	nil	-	+	nil	nil
g'	velar	stop	palatalized	-	+	nil	nil
g:	velar	stop	nil	-	+	nil	nil
h	glottal	fricatives	nil	-	+	nil	nil
h:	glottal	fricatives	nil	-	+	nil	nil
h	glottal	fricatives	nil	-	+	nil	nil
j	palatal	glides	nil	+	+	nil	nil
j:	palatal	glides	nil	+	+	nil	nil
j_r							
k	velar	stop	nil	-	-	nil	nil
k'	velar	stop	palatalized	-	-	nil	nil
k:	velar	stop	nil	-	-	nil	nil
l	prealveolar	lateral	nil	+	+	nil	nil
l'	prealveolar	lateral	palatalized	+	+	nil	nil
l:	prealveolar	lateral	nil	+	+	nil	nil
l=							
l=:							
m	bilabial	nasals	nil	+	+	nil	nil
m:	bilabial	nasals	nil	+	+	nil	nil
n	prealveolar	nasals	nil	+	+	nil	nil
n:	prealveolar	nasals	nil	+	+	nil	nil
p	bilabial	stop	nil	-	-	nil	nil
p'	bilabial	stop	palatalized	-	-	nil	nil
p:	bilabial	stop	nil	-	-	nil	nil
r	prealveolar	trills	nil	+	+	nil	nil
r'	prealveolar	trills	palatalized	+	+	nil	nil
r:	prealveolar	trills	nil	+	+	nil	nil
r=							
r=:							
r_r	prealveolar	trills	nil	-	+	nil	nil
s	prealveolar	fricatives	nil	-	-	nil	nil
s:	prealveolar	fricatives	nil	-	-	nil	nil
s	alveolopalatal	fricatives	nil	-	-	nil	nil
t	prealveolar	stop	nil	-	-	nil	nil
t:	prealveolar	stop	nil	-	-	nil	nil
tS	postalveolar	affricates	nil	-	-	nil	nil
tS'	postalveolar	affricates	palatalized	-	-	nil	nil
tS:	postalveolar	affricates	nil	-	-	nil	nil
ts	prealveolar	affricates	nil	-	-	nil	nil
ts:	prealveolar	affricates	nil	-	-	nil	nil
ts	alveolopalatal	affricates	nil	-	-	nil	nil
tz	alveolopalatal	affricates	nil	-	+	nil	nil

Phones	Place_con	Manner_con	palatalization	Sonor	Voicing	Manner_vow	Place_vow
v	labiodental	fricatives	nil	-	+	nil	nil
v’	labiodental	fricatives	palatalized	-	+	nil	nil
v:	labiodental	fricatives	nil	-	+	nil	nil
w	labiodental	glides	nil	+	+	nil	nil
x	velar	fricatives	nil	-	-	nil	nil
z	prealveolar	fricatives	nil	-	+	nil	nil
z:	prealveolar	fricatives	nil	-	+	nil	nil
z	alveolopalatal	fricatives	nil	-	+	nil	nil
”1	nil	nil	nil	nil	+	close	central
”a	nil	nil	nil	nil	+	open	central
”e	nil	nil	nil	nil	+	close-mid	front
”i	nil	nil	nil	nil	+	close	front
”o	nil	nil	nil	nil	+	close-mid	back
”u	nil	nil	nil	nil	+	close	back
1	nil	nil	nil	nil	+	close	central
2	nil	nil	nil	nil	+	close-mid	front
2:	nil	nil	nil	nil	+	close-mid	front
A	nil	nil	nil	nil	+	open	back
E	nil	nil	nil	nil	+	open-mid	front
E:	nil	nil	nil	nil	+	open-mid	front
I	nil	nil	nil	nil	+	close-mid	front
a	nil	nil	nil	nil	+	open	central
a:	nil	nil	nil	nil	+	open	central
e	nil	nil	nil	nil	+	close-mid	front
e:	nil	nil	nil	nil	+	close-mid	front
e	nil	nil	nil	nil	+	close-mid	front
i	nil	nil	nil	nil	+	close	front
i:	nil	nil	nil	nil	+	close	front
o	nil	nil	nil	nil	+	close-mid	back
o:	nil	nil	nil	nil	+	close-mid	back
o	nil	nil	nil	nil	+	close-mid	back
u	nil	nil	nil	nil	+	close	back
u:	nil	nil	nil	nil	+	close	back
y	nil	nil	nil	nil	+	close	front
y:	nil	nil	nil	nil	+	close	front
Eu	nil	nil	nil	nil	+	open-mid	front
au	nil	nil	nil	nil	+	open	central
i_â	nil	nil	nil	nil	+	close	front
i_ê	nil	nil	nil	nil	+	close	front
i_û	nil	nil	nil	nil	+	close	front
ou	nil	nil	nil	nil	+	close-mid	back
u_ô	nil	nil	nil	nil	+	close	back

Table A.1: Summary of articulatory features

Bibliography

[1] B. Abraham, S. Umesh, and N. M. Joy. “Joint Estimation of Articulatory Features and Acoustic Models for Low-Resource Languages”. In: *INTERSPEECH*. 2017.

[2] D. Amodei and et al. “Deep Speech 2: End-to-End Speech Recognition in English and Mandarin”. In: *CoRR* abs/1512.02595 (2015). URL: `http://arxiv.org/abs/1512.02595`.

[3] Arcadian. “Head and Neck Overview (from http://training.seer.cancer.gov/head-neck/anatomy/overview.html)”. Source http://training.seer.cancer.gov/head-neck /anatomy/overview.html. 15 February 2007.

[4] “Automatic Speech recognition In Reverberant Environ-ments (ASpIRE) Challenge”. In: vol. 2015. 2015. URL: `http://www.iarpa.gov/index.php/working-with-iarpa/prize-challenges/306-automatic-speech-in-reverberant-environments-aspire-challenge`.

[5] J. Barker et al. “The fifth ’CHiME’ Speech Separation and Recognition Challenge: Dataset, task and baselines”. In: (Mar. 2018).

[6] C. Barras, L. Lamel, and J.-L. Gauvain. “Automatic Transcription of Compressed Broadcast Audio”. In: *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing.* Salt Lake City, USA, 2001, pp. 265–268.

[7] Y. C. Barry, C. Shuangyu, and S. Sunil. “Learning discriminative temporal patterns in speech: development of novel TRAPS-like classifiers”. In: *INTERSPEECH*. 2003.

[8] L. Besacier et al. “Automatic speech recognition for under-resourced languages: A survey”. In: *Speech Communication* 56.0 (2014), pp. 85 –100. ISSN: 0167-6393. DOI: `http://dx.doi.org/10.1016/j.specom.2013.07.008`. URL: `http://www.sciencedirect.com/science/article/pii/S0167639313000988`.

[9] N. Bhattacharyya. “The prevalence of voice problems among adults in the United States.” In: *Laryngoscope* 124.10 (2014), pp. 2359–2362.

[10] P. Boersma and D. Weenink. *Praat: doing phonetics by computer (Version 6.1.01).* 2009. URL: `http://www.praat.org`.

[11] B. Bollepalli, A. W. Black, and K. Prahallad. In: *INTERSPEECH*. ISCA, 2012.

[12] M. Borsky, P. Mizera, and P. Pollak. “Noise and Channel Normalized Cepstral Features for Far-Speech Recognition”. In: *Speech and Computer. Cham: Springer International Publishing AG.* Vol. 3. 2014, pp. 241–248.

[13] W. Byrne and et al. “Automatic Recognition of Spontaneous Speech for Access to Multilingual Oral History Archives”. In: *IEEE Trans. on Speech and Audio Processing* Vol.12.No.4 (2004), pp. 420–435.

[14] O. Cetin et al. “Monolingual and crosslingual comparison of tandem features derived from articulatory and phone MLPS”. In: *Automatic Speech Recognition Understanding, 2007. ASRU. IEEE Workshop on.* 2007, pp. 36–41.

[15] J. Chaloupka et al. “Voice Technology Applied for Building a Prototype Smart Room”. In: *Voice Technology Applied for Building a Prototype Smart Room. Springer Series LNCS.* Vol. 5398. 2009, pp. 104–111.

[16] W. Chan et al. “Listen, Attend and Spell”. In: *ArXiv* abs/1508.01211 (2015).

[17] S. Chang, M. Wester, and S. Greenberg. “An elitist approach to automatic articulatory-acoustic feature classification for phonetic characterization of spoken language”. In: *Speech Communication* 47.3 (2009), pp. 290–311.

[18] Y. L. Chow. “Maximum mutual information estimation of HMM parameters for continuous speech recognition using the N-best algorithm”. In: *International Conference on Acoustics, Speech, and Signal Processing.* 1990, 701–704 vol.2.

[19] R. Cmejla et al. “Bayesian Changepoint Detection for the Automatic Assessment of Fluency and Articulatory Disorders”. In: *Speech Commun.* 55.1 (Jan. 2013), pp. 178–189. ISSN: 0167-6393. DOI: 10.1016/j.specom.2012.08.003. URL: http://dx.doi.org/10.1016/j.specom.2012.08.003.

[20] J. Cui et al. “Recent Improvements in Neural Network Acoustic Modeling for LVCSR in Low Resource Languages”. In: *Proc. of Interspeech 2014: 15th Annual Conference of the Interantional Speech Communication Association.* Singapore, 2014.

[21] G. E. Dahl, D. Li Y. Dong, and A. Acero. “Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition”. In: *Audio, Speech, and Language Processing, IEEE Transactions on* 20.1 (2012), pp. 30–42.

[22] E.E. David and O. G. Selfridge. “Eyes and Ears for Computers”. In: *Proceedings of the IRE* 50.5 (1962), pp. 1093–1101. ISSN: 0096-8390.

[23] L. Deng, G. Hinton, and B. Kingsbury. “New types of deep neural network learning for speech recognition and related applications: an overview”. In: *Acoustics, Speech and Signal Processing (ICASSP).* 2013, pp. 8599–8603.

[24] J. Droppo and A. Acero. “Environmental Robustness”. In: *Springer Handbook of Speech Processing.* Springer, 2008, pp. 653–680.

[25] M. Ernestus, L. Kočková-Amortová, and P. Pollák. “The Nijmegen Corpus of Casual Czech”. In: *Proc. of LREC 2014: 9th International Conference on Language Resources and Evaluation.* Reykjavik, Iceland, 2014, pp. 365–370.

[26] E.Yilmaz et al. “Articulatory Features for ASR of Pathological Speech”. In: *CoRR* abs/1807.10948 (2018). arXiv: 1807.10948. URL: http://arxiv.org/abs/1807.10948.

[27] R. Fér et al. “Multilingually Trained Bottleneck Features in Spoken Language Recognition”. In: *Computer Speech and Language* 2017.46 (2017), pp. 252–267. ISSN: 0885-2308. DOI: 10.1016/j.csl.2017.06.008. URL: https://www.fit.vut.cz/research/publication/11518.

[28] J. Fiala. “DNN-HMM Based Multilingual Recognizer of Telephone Speech, Master Thesis”. In: Czech Technical University in Prague, 2016.

[29] P. Fousek. "Extraction of Features for Automatic Recognition of Speech Based on Spectral Dynamics, PhD Thesis". In: Czech Technical University in Prague, 2007.

[30] P. Fousek, P. Mizera, and P. Pollak. "CtuCopy feature extraction tool". Available at `http://noel.feld.cvut.cz/speechlab/`.

[31] J. Frankel and S. King. "A Hybrid ANN/DBN Approach to Articulatory Feature Recognition". In: *Proc. Eurospeech.* Lisbon, 2005.

[32] J. Frankel, M. Wester, and S. King. "Articulatory feature recognition using dynamic Bayesian networks". In: *Computer Speech & Language* 21.4 (2007), pp. 620–640.

[33] J. Frankel et al. "Articulatory feature classifiers trained on 2000 hours of telephone speech". In: *Proceedings of Interspeech.* Antwerp, Belgium, 2007.

[34] M. J. F. Gales. "Factored Semi-tied Covariance Matrices". In: *Proceedings of the 13th International Conference on Neural Information Processing Systems.* NIPS'00. Denver, CO: MIT Press, 2000, pp. 749–755. URL: `http://dl.acm.org/citation.cfm?id=3008751.3008860`.

[35] T. Gan, W. Menzel, and J. Zhang. "Using the Tandem Approach for AF Classification in an AVSR System". In: *Proceedings of the 5th International Symposium on Neural Networks: Advances in Neural Networks, Part II.* ISNN '08. Beijing, China, 2008, pp. 830–839. ISBN: 978-3-540-87733-2.

[36] P. Ghahremani et al. "A pitch extraction algorithm tuned for automatic speech recognition". In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 2014, pp. 2494–2498. DOI: `10.1109/ICASSP.2014.6854049`.

[37] J. T. Goodman. "A bit of progress in language modeling". In: *Computer Speech & Language* 15.4 (2001), pp. 403 –434. ISSN: 0885-2308. DOI: `https://doi.org/10.1006/csla.2001.0174`. URL: `http://www.sciencedirect.com/science/article/pii/S0885230801901743`.

[38] F. Grezl. "Trap-based Probabilistic Features for Automatic Speech Recognition, PhD Thesis". In: Brno University of Technology, 2009.

[39] F. Grezl and P. Fousek. "Optimizing bottle-neck features for lvcsr". In: *2008 IEEE International Conference on Acoustics, Speech and Signal Processing.* 2008, pp. 4729–4732. DOI: `10.1109/ICASSP.2008.4518713`.

[40] F. Grezl et al. "Probabilistic and Bottle-Neck Features for LVCSR of Meetings". In: *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07.* Vol. 4. 2007, pp. IV–757–IV–760. DOI: `10.1109/ICASSP.2007.367023`.

[41] H. Guangpu. "Articulatory Phonetic Features for Robust Speech Recognition". In: *Ph.D. Thesis, Nanyang Technological University, School of Electrical & Electronic Engineering* (2012).

[42] R. Gupta et al. "Pathological speech processing: State-of-the-art, current challenges, and future directions". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 2016, pp. 6470–6474. DOI: `10.1109/ICASSP.2016.7472923`.

[43] R. H*a*eb-Umbach and H. Ney. “Linear discriminant analysis for improved large vocabulary continuous speech recognition”. In: *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing.* Vol. 1. 1992, 13–16 vol.1. DOI: `10.1109/ICASSP.1992.225984`.

[44] A. Hanulíková and S. Hamann. “Illustrations of the IPA: Slovak”. In: *Journal of the International Phonetic Association* 40 (Dec. 2010), pp. 373 –378. DOI: `10.1017/S0025100310000162`.

[45] M. Hasegawa-Johnson et al. “Landmark-Based Speech Recognition: Report of the 2004 Johns Hopkins Summer Workshop”. In: *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP ’05). IEEE International Conference on.* Vol. 1. 2005, pp. 213–216.

[46] H. Hermansky. “Perceptual Linear Predictive (PLP) Analysis of Speech”. In: *J. Acoust. Soc. Am.* 57.4 (Apr. 1990), pp. 1738–52.

[47] H. Hermansky, D.P.W. Ellis, and S. Sharma. “Tandem connectionist feature extraction for conventional HMM systems”. In: *Acoustics, Speech, and Signal Processing, 2000. ICASSP ’00. Proceedings. 2000 IEEE International Conference on.* Vol. 3. 2000, 1635–1638 vol.3.

[48] G. Hinton and et al. “Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups”. In: *Signal Processing Magazine, IEEE* 29.6 (2012), pp. 82–97.

[49] J. P. Hosom. “Automatic phoneme alignment based on acoustic-phonetic modeling”. In: *In ICSLP.* 2002, pp. 357–360.

[50] Institute of the Czech National Corpus. *Corpus ORAL 2006 and ORAL 2008 and ORAL 2013, Institute of the Czech National Corpus FF UK.* `http://www.korpus.cz`. Prague.

[51] Institute of the Czech National Corpus. *SYN2006PUB Corpus.* `http://ucnk.ff.cuni.cz/english/syn2006pub.php`. Prague, 2006.

[52] P. Ircing et al. “On large vocabulary continuous speech recognition of highly inflectional language - Czech”. In: *INTERSPEECH.* 2001, pp. 487–490.

[53] W. Jassem. “Polish”. In: *Journal of the International Phonetic Association* 33.1 (2003), pp. 103–107. ISSN: 00251003, 14753502. URL: `http://www.jstor.org/stable/44526910`.

[54] F. Jelinek, L. Bahl, and R. Mercer. “Design of a linguistic statistical decoder for the recognition of continuous speech”. In: *Information Theory, IEEE Transactions on* 21.3 (1975), pp. 250–256.

[55] H. Jiang. “Discriminative training of HMMs for automatic speech recognition: A survey”. In: *Computer Speech Language* 24.4 (2010), pp. 589 –608. ISSN: 0885-2308. DOI: `https://doi.org/10.1016/j.csl.2009.08.002`. URL: `http://www.sciencedirect.com/science/article/pii/S0885230809000606`.

[56] Y. Jiao, V. Berisha, and J. Liss. “Interpretable Phonological Features for Clinical Applications”. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 2017, pp. 5045–5049.

[57] D. Johnson and et al. *ICSI Quicknet Software Package.* 2004.

[58] A. Kahn and I. Steiner. “Qualitative Evaluation and Error Analysis of Phonetic Segmentation”. In: *28. Konferenz Elektronische Sprachsignalverarbeitung.* Saarbrücken, Germany, 2017, pp. 138–144.

[59] I. Karaulov and D. Tkanov. “Attention model for articulatory features detection”. In: *ArXiv* abs/1907.01914 (2019). URL: `https://arxiv.org/abs/1907.01914`.

[60] S. King and P. Taylor. “Detection of phonological features in continuous speech using neural networks”. In: *Computer Speech & Language* 14.4 (2000), pp. 333–353.

[61] S. King et al. “Speech production knowledge in automatic speech recognition”. In: *J. Acoust. Soc. Amer.* 121.2 (2007), pp. 723–742.

[62] B. E. D. Kingsbury. *Perceptually Inspired Signal-processing Strategies for Robust Speech Recognition in Reverberant Environments.* Tech. rep. PhD Dissertation, 1998.

[63] B. E. D. Kingsbury, N. Morgan, and S. Greenberg. “Robust speech recognition using the modulation spectrogram”. In: *Speech Communication* 25.1-3 (1998), pp. 117–132.

[64] K. Kinoshita et al. “The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech”. In: Oct. 2013. DOI: `10.1109/WASPAA.2013.6701894`.

[65] K. Kirchhoff. “Robust speech recognition using articulatory information”. PhD thesis. Der Technischen Fakult*ae*t der Universit*ae*t Bielefeld, 1999.

[66] K. Kirchhoff, G. A. Fink, and G. Sagerer. “Combining Acoustic and Articulatory Feature Information For Robust Speech Recognition”. In: *Speech Communication* 37.3-4 (2002), pp. 303–319.

[67] T. Ko et al. “Audio augmentation for speech recognition.” In: *INTERSPEECH.* ISCA, 2015, pp. 3586–3589. URL: `http://dblp.uni-trier.de/db/conf/interspeech/interspeech2015.html#KoPPK15`.

[68] D. Kocharov et al. “Articulatory Motivated Acoustic Features for Speech Recognition”. In: *Interspeech.* Sept. 2005, pp. 1101–1104.

[69] A. Kolman and P. Pollak. “Speech reduction in Czech”. In: *Proc. of LabPhone 14, The 14th Conference on Laboratory Phonology.* Tokyo, Japan, 2014.

[70] M. Korvas et al. “Free English and Czech Telephone Speech Corpus Shared Under the CC-BY-SA 3.0 License”. In: *Proc. of LREC 2014: 9th International Conference on Language Resources and Evaluation.* Reykjavik, Iceland, 2014, pp. 365–370.

[71] K. Kumar, Ch. Kim, and R. M. Stern. “Delta-spectral cepstral coefficients for robust speech recognition.” In: *ICASSP.* IEEE, 2011, pp. 4784–4787. ISBN: 978-1-4577-0539-7.

[72] P. Lal. “Cross-lingual Automatic Speech Recognition using Tandem Features”. In: *Ph.D. Thesis, University of Edinburgh, Institute for Language, Cognition and Computation* (2011).

[73] P. Lal and S. King. “Cross-lingual Automatic Speech Recognition using Tandem Features”. In: *IEEE Transactions on Audio, Speech, and Language Processing* To appear (2013). ISSN: 1558-7916. DOI: `10.1109/TASL.2013.2277932`.

[74] K. Lee and H. Hon. “Speaker-independent phone recognition using hidden markov models”. In: *IEEE Transactions on Audio, Speech & Language Processing* 37.11 (1989), pp. 1641–1648.

[75] M. Lehr, K. Gorman, and I. Shafran. “Discriminative pronunciation modeling for dialectal speech recognition”. In: *Proc. of Interspeech 2014: 15th Annual Conference of the Interantional Speech Communication Association*. Singapore, 2014, pp. 1458–1462.

[76] J. Li et al. *Robust Automatic Speech Recognition - A Bridge to Practical Applications (1st Edition)*. Elsevier, 2015. URL: `https://www.microsoft.com/en-us/research/publication/robust-automatic-speech-recognition-a-bridge-to-practical-applications-1st-edition-306-pages/`.

[77] K. Livescu. “Feature-Based Pronunciation Modeling for Automatic Speech Recognition”. In: *Ph.D. Thesis, MIT Department of Electrical Engineering and Computer Science* (2005).

[78] K. Livescu and et al. “Articulatory Feature-Based Methods for Acoustic and Audio-Visual Speech Recognition: Summary from the 2006 JHU Summer workshop”. In: *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. Vol. 4. 2007, pp. IV–621–IV–624.

[79] K. Livescu and et al. “Manual Transcription of Conversational Speech at the Articulatory Feature Level”. In: *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. Vol. 4. 2007, pp. IV–953–IV–956.

[80] K. Livescu, E. Fosler-Lussier, and F. Metze. “Subword Modeling for Automatic Speech Recognition: Past, Present, and Emerging Approaches”. In: *Signal Processing Magazine, IEEE* 29.6 (2012), pp. 44–57.

[81] A. Lozano-Diez et al. “Analysis and Optimization of Bottleneck Features for Speaker Recognition”. In: *Odyssey 2016*. 2016, pp. 352–357. DOI: `10.21437/Odyssey.2016-51`. URL: `http://dx.doi.org/10.21437/Odyssey.2016-51`.

[82] J. Matousek and M. Klima. “Automatic Phonetic Segmentation Using the KALDI Toolkit”. In: *Text, Speech and Dialogue. TSD 2017. Lecture Notes in Computer Science, vol 10415*. Springer, Berlin, Heidelberg, 2017, pp. 138–146.

[83] J. Matoušek, D. Tihelka, and J. Psutka. “Experiments with Automatic Segmentation for Czech Speech Synthesis”. In: *Text, Speech and Dialogue. TSD 2003*. Springer, Berlin, Heidelberg, 2013, pp. 287–294.

[84] Newton Media. “Newton Dictate Home page”. http://www.diktovani.cz. 2013.

[85] F. Metze. “Articulatory features for conversational speech recognition, PhD Thesis”. In: (Jan. 2005).

[86] V. Mitra et al. “Articulatory Information for Noise Robust Speech Recognition”. In: *Audio, Speech, and Language Processing, IEEE Transactions on* 19.7 (2011), pp. 1913–1924.

[87] Vikramjit Mitra. “ARTICULATORY INFORMATION FOR ROBUST SPEECH RECOGNITION, PhD Thesis”. In: (Jan. 2010).

[88] P. Mizera. “The contribution of differential cepstral coefficients within AF estimation”. In: *Proceedings of 18th International Student Conference on Electrical Engineering, Prague: Czech Technical University*. Vol. 3. 2014, 1635–1638 vol.3.

[89] P. Mizera et al. “Impact of Irregular Pronunciations for Phonetic Segmentation of Nijmegen Corpus of Casual Czech”. In: *Text, Speech and Dialogue. TSD 2014*. Springer, Berlin, Heidelberg, 2014, pp. 499–506.

[90] M. Mohri, F. C. N. Pereira, and M. Riley. “Speech Recognition with Weighted Finite-State Transducers”. In: *Handbook on Speech Processing and Speech Communication, Part E: Speech recognition.* 2007. URL: `http://www.cs.nyu.edu/~mohri/postscript/hbka.pdf`.

[91] N. Morgan and H. Bourlard. “Continuous speech recognition: An introduction to the hybrid HMM/connectionist approach”. In: *Signal Processing Magazine, IEEE* 12.3 (1995), pp. 24–42.

[92] N. Morgan and et al. “Pushing the envelope - aside [speech recognition]”. In: *Signal Processing Magazine, IEEE* 22.5 (2005), pp. 81–88. ISSN: 1053-5888. DOI: `10.1109/MSP.2005.1511826`.

[93] A. B. N*ae*ss, K. Livescu, and R. Prabhavalkar. “Articulatory Feature Classification Using Nearest Neighbors.” In: *INTERSPEECH.* Florence, Italy: ISCA, 2011, pp. 2301–2304.

[94] J. Nouza and J. Silovský. “Adpating Lexical and Language Models for Transcription of Highly Spontaneous Spoken Czech”. In: *Proc. of Text, Speech, and Dialogue, LNAI Vol. 6231.* Brno, Czech Republic, 2010, pp. 377–384.

[95] J. Nouza, J. Ždánský, and P. Červa. “System for automatic collection, annotation and indexing of Czech broadcast speech with full-text search”. In: *Proc. of 15th IEEE MELECON Conference.* La Valleta, Malta, 2010, pp. 202–205.

[96] J. Nouza, J. Zdansky, and P. David. “Fully Automated Approach to Broadcast News Transcription in Czech Language”. In: *Proc. of TSD 2004 Conference, Springer Series LNCS.* Vol. 3206. 2004, pp. 401–408.

[97] J. Nouza et al. “System for producing subtitles to internet audio-visual documents”. In: *38th International Conference on Telecommunications and Signal Processing, TSP 2015, Prague, Czech Republic, July 9-11, 2015.* 2015, pp. 1–5.

[98] M. Novotný et al. “Automatic Evaluation of Articulatory Disorders in Parkinson’s Disease”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.9 (2014), pp. 1366–1378. ISSN: 2329-9290.

[99] O. Abdel-Hamid and et al. “Applying Convolutional Neural Networks concepts to hybrid NN-HMM model for speech recognition”. In: *Acoustics, Speech and Signal Processing (ICASSP).* 2012, pp. 4277–4280.

[100] Z. Palková. *Czech phonetics and phonology.* In Czech language: *Fonetika a fonologie češtiny.* Karolinum, 1997.

[101] V. Panayotov et al. “Librispeech: An ASR corpus based on public domain audio books”. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 2015, pp. 5206–5210.

[102] J. P. Pinto et al. “Exploiting contextual information for improved phoneme recognition.” In: *ICASSP.* IEEE, Feb. 10, 2010, pp. 4449–4452.

[103] J. P. Pinto et al. *Significance of Contextual Information in Phoneme Recognition.* Idiap-RR Idiap-RR-28-2007. IDIAP, 2007.

[104] P. Pollak and J. Cernocky. *Czech SPEECON Adult Database.* Tech. rep. 2004.

[105] P. Pollák and V. Hanzl. "Tool for Czech Pronunciation Generation Combining Fixed Rules with Pronunciation Lexicon and Lexicon Management Tool". In: *LREC.* 2002.

[106] P. Pollak et al. "SpeechDat(E) - Eastern European Telephone Speech Databases". In: May 2000, pp. 20–25.

[107] D. Povey and et al. "The Kaldi Speech Recognition Toolkit". In: *Proc of ASRU 2011, IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.* Big Island, Hawaii, US, 2011.

[108] D. Povey, A. Ghoshal, and et al. "The Kaldi Speech Recognition Toolkit". In: *Proc of ASRU 2011, IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.* Big Island, Hawaii, US, 2011.

[109] D. Povey et al. "Generating exact lattices in the WFST framework". In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 2012, pp. 4213–4216. DOI: 10.1109/ICASSP.2012.6288848.

[110] D. Povey et al. "Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI". In: *INTERSPEECH.* 2016.

[111] R. Prabhavalkar et al. "Discriminative articulatory models for spoken term detection in low-resource conversational settings". In: *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on.* 2013, pp. 8287–8291.

[112] J. Pratibha, H. Hermansky, and B. Kingsbury. "Distributed speech recognition using noise-robust MFCC and traps-estimated manner features". In: *INTERSPEECH.* Denver, Colorado, 2002, pp. 487–490.

[113] V. Prochazka et al. "Performance of Czech Speech Recognition with Language Models Created from Public Resources". In: *Radioengineering* 20 (2011), pp. 1002–1008.

[114] V. Prochazka et al. "Performance of Czech Speech Recognition with Language Models Created from Public Resources". In: *Radioengineering* 20.4 (2011), pp. 1002–1008.

[115] J. Psutka et al. *Mluvíme s počítačem česky.* Prague: Academia, 2006, p. 752. ISBN: 80-200-1309-1. URL: http://www.kky.zcu.cz/en/publications/PsutkaJ_2006_Mluvimes.

[116] J. Psutka et al. "Recognition of spontaneously pronounced TV ice-hockey commentary". In: *Proc. of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition.* Tokyo, 2003, pp. 83–86.

[117] J. Rajnoha and P. Pollak. "ASR systems in Noisy Environment: Analysis and Solutions for Increasing Noise Robustness". In: *Radioengineering* 20.1 (2011).

[118] J. Rajnoha and P. Pollák. "Czech Spontaneous Speech Collection and Annotation: The Database of Technical Lectures". In: *Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions, LNCS Vol. 5641.* Prague, Czech Republic, 2009.

[119] R. Rasipuram and M. Magimai-Doss. "Integrating articulatory features using Kullback-Leibler divergence based acoustic model for phoneme recognition". In: *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on.* 2011, pp. 5192–5195. DOI: `10.1109/ICASSP.2011.5947527`.

[120] R. Rasipuram and M. Magimai-Doss. *Multitask Learning to Improve Articulatory Feature Estimation and Phoneme Recognition.* Idiap-RR Idiap-RR-21-2011. Idiap, June 2011.

[121] A. Rendel et al. "Toward Automatic Phonetic Segmentation for TTS". In: *Acoustics, Speech and Signal Processing (ICASSP).* Kyoto, Japan, 2012, pp. 4533–4536.

[122] P. Roach. "English phonetics and phonology: a practical course. Cambridge: Cambridge University Press, 1983. Pp. x, 212". In: *RELC Journal* 15.1 (1984), pp. 117–118. DOI: `10.1177/003368828401500113`. eprint: `https://doi.org/10.1177/003368828401500113`. URL: `https://doi.org/10.1177/003368828401500113`.

[123] F. Rudzicz. "Articulatory Knowledge in the Recognition of Dysarthric Speech". In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.4 (2011), pp. 947–960. ISSN: 1558-7916. DOI: `10.1109/TASL.2010.2072499`.

[124] T. N. Sainath et al. "Making Deep Belief Networks effective for large vocabulary continuous speech recognition". In: *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on.* 2011, pp. 30–35.

[125] H. Sakoe and S. Chiba. "Dynamic programming algorithm optimization for spoken word recognition". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 26.1 (1978), pp. 43–49. ISSN: 0096-3518. DOI: `10.1109/TASSP.1978.1163055`.

[126] B. Schuppler. "Rethinking classification results based on read speech, or: why improvements do not always transfer to other speaking styles". In: *International Journal of Speech Technology* 20.3 (2017), pp. 699–713. ISSN: 1572-8110. DOI: `10.1007/s10772-017-9436-y`. URL: `https://doi.org/10.1007/s10772-017-9436-y`.

[127] B. Schuppler, M. Adda-Decker, and J. A. Morales-Cordovilla. "Pronunciation variation in read and conversational Austrian German". In: *Proc. of Interspeech 2014: 15th Annual Conference of the Interantional Speech Communication Association.* Singapore, 2014.

[128] P. Schwarz. "Phoneme Recognition based on Long Temporal Context, PhD Thesis". In: Brno University of Technology, 2009.

[129] L. M. Seltzer, Y. Dong, and W. Yongqiang. "An investigation of deep neural networks for noise robust speech recognition". In: *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, Canada, 2013.*

[130] Z. Shi-Xiong, M. W. Mak, and H.M. Meng. "Speaker Verification via High-Level Feature Based Phonetic-Class Pronunciation Modeling". In: *Computers, IEEE Transactions on* 56.9 (2007), pp. 1189–1198.

[131] A. Stolcke et al. "Highly accurate phonetic segmentation using boundary correction models and system fusion." In: *Proc. of ICASSP.* Florence, Italy, 2014.

[132] Andreas Stolcke. "SRILM – An extensible language modeling toolkit". In: *IN PROCEEDINGS OF THE 7TH INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING (ICSLP 2002.* 2002, pp. 901–904.

[133] T. Szende. “Illustrations of the IPA: Hungarian”. In: *Journal of the International Phonetic Association* 24 (1994), pp. 91–94.

[134] D. T. Toledano, L. A. H. Gómez, and L. V. Grande. “Automatic phoneme segmentation”. In: *IEEE Transactions on Speech and Audio Processing* 11.6 (2003), pp. 617–625.

[135] F. Torreira, M. Adda-Decker, and M. Ernestus. “The Nijmegen Corpus of Casual French”. In: *Speech Communication* 52 (2010), pp. 201–221.

[136] J. Trmal and et al. “A keyword search system using open source software”. In: *2014 IEEE Workshop on Spoken Language Technology, SLT 2014 - Proceedings* (Apr. 2015), pp. 530–535. DOI: 10.1109/SLT.2014.7078630.

[137] F. Valente, M. M. Doss, and W. Wang. “Analysis and Comparison of Recent MLP Features for LVCSR Systems”. In: *Proc. of Interspeech.* Florence, Italy, 2011.

[138] J. Vanek and J. Psutka. “Gender-Dependent Acoustic Models Fusion Developed for Automatic Subtitling of Parliament Meetings Broadcasted by the Czech TV”. In: *Proc. of Text, Speech and Dialog.* Brno, Czech Republic, 2010, pp. 431–438.

[139] J. Černocký. “Temporal processing for feature extraction in speech recognition, shortened version of habilitation thesis”. In: *Vědecké spisy VUT.* Edice Habilitační a inaugurační spisy, sv. 112. Brno, CZ: Publishing house of Brno University of Technology VUTIUM, 2003, pp. 1–30. ISBN: 80-214-2395-1. URL: https://www.fit.vut.cz/research/publication/7240.

[140] K. Vesely. “Semi-Supervised Training of Deep Neural Networks for Speech Recognition, PhD Thesis”. In: Brno University of Technology, 2017.

[141] K. Vesely, L. Burget, and F. Grezl. “Parallel Training of Neural Networks for Speech Recognition”. In: *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010).* Vol. 2010. 9. Makuhari, Chiba, JP: International Speech Communication Association, 2010, pp. 2934–2937. URL: http://www.fit.vutbr.cz/research/view_pub.php?id=9364.

[142] J. Volín. “Grammar of Contemporary Czech”. In: Cvrček, V. and et al. In Czech language: Mluvnice současné češtiny. Karolinum, 2013. Chap. Phonetic and Phonology, pp. 35–64.

[143] J. C. Wells et al. “Czech SAMPA Home Page”. http://www.phon.ucl.ac.uk/home/sampa/home.htm, 2003.

[144] K. H. Wong et al. “Exploring articulatory characteristics of Cantonese dysarthric speech using distinctive features”. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 2016, pp. 6495–6499. DOI: 10.1109/ICASSP.2016.7472928.

[145] H. Xu et al. “Neural Network Language Modeling with Letter-Based Features and Importance Sampling”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018), pp. 6109–6113.

[146] Q. Yanmin and L. Jia. “Articulatory Feature based Multilingual MLPs for Low-Resource Speech Recognition.” In: *INTERSPEECH.* ISCA, 2012.

[147] I. Yanushevskaya and D. Bunčic. “Illustrations of the IPA: Russian”. In: *Journal of the International Phonetic Association* 45 (Aug. 2015), pp. 221–228. DOI: 10.1017/S0025100314000395.

[148] D. Yu et al. “Boosting attribute and phone estimation accuracies with deep neural networks for detection-based speech recognition”. In: *Acoustics, Speech and Signal Processing (ICASSP)*. 2012, pp. 4169–4172.

List of candidate's work related to the thesis

The percentage is even for all listed authors at each publication.

Journals (Impact)

- M. Borský, P. Mizera, P. Pollák, and J. Nouza. "Dithering techniques in automatic recognition of speech corrupted by MP3 compression: Analysis, solutions and experiments". English. In: *Speech Communication* 86 (2017), pp. 75–84. ISSN: 0167-6393. DOI: 10.1016/j.specom.2016.11.007. URL: http://www.sciencedirect.com/science/article/pii/S0167639316300863

- M. Borský, P. Pollák, and P. Mizera. "Advanced Acoustic Modelling Techniques in MP3 Speech Recognition". English. In: *EURASIP Journal on Audio Speech and Music Processing* 2015:20 (2015). ISSN: 1687-4722. DOI: 10.1186/s13636-015-0064-7. URL: http://asmp.eurasipjournals.com/content/2015/1/20

- P. Mizera and P. Pollák. "Robust Neural Network-Based Estimation of Articulatory Features for Czech". English. In: *Neural Network World* 24.5 (Oct. 2014), pp. 463–478. ISSN: 1210-0552. DOI: 10.14311/NNW.2014.24.027. URL: http://nnw.cz/obsahy14.html\#5-2014

Conferences

- P. Mizera and P. Pollák. "Automatic Phonetic Segmentation and Pronunciation Detection with Various Approaches of Acoustic Modeling". English. In: *Speech and Computer.* Vol. 11096. LNAI. Leipzig, GE, 2018. ISBN: 978-3-319-99578-6. DOI: 10.1007/978-3-319-99579-3_44

- P. Mizera and P. Pollák. "Improving of LVCSR for Casual Czech Using Publicly Available Language Resources". English. In: *Speech and Computer.* Vol. LNAI 10458. Lecture Notes in Artificial Intelligence. Hatfield, UK, 2017, pp. 427–437. ISBN: 978-3-319-66428-6. DOI: 10.1007/978-3-319-66429-3_42. URL: https://link.springer.com/chapter/10.1007/978-3-319-66429-3_42

- P. Mizera, J. Fiala, A. Brich, and P. Pollák. "KALDI Recipes for the Czech Speech Recognition Under Various Conditions". English. In: *Text, Speech, and Dialogue. 19th International Conference, TSD 2016.* Lecture Notes in Artificial Intelligence. Brno, CR, 2016. ISBN: 978-3-319-45510-5. DOI: 10.1007/978-3-319-45510-5_45

- Z. Patč, P. Mizera, and P. Pollák. "Phonetic Segmentation Using KALDI and Reduced Pronunciation Detection in Causal Czech Speech". English. In: *Text, Speech, and Dialogue. 18th International Conference, TSD 2015.* Lecture Notes in Artificial Intelligence. Pilsen, CR, 2015, pp. 433–441. ISBN: 978-3-319-24032-9. DOI: 10.1007/978-3-319-24033-6_49. URL: http://link.springer.com/chapter/10.1007/978-3-319-24033-6_49

- P. Mizera and P. Pollák. "Improved Estimation of Articulatory Features Based on Acoustic Features with Temporal Context". English. In: *Text, Speech, and Dialogue. 18th International Conference, TSD 2015.* Lecture Notes in Artificial Intelligence. Pilsen, CR, 2015, pp. 560–568. ISBN: 978-3-319-24032-9. DOI: 10.1007/978-3-

319-24033-6_63. URL: http://link.springer.com/chapter/10.1007/978-3-319-24033-6_63

- M. Borský, P. Mizera, and P. Pollák. “Spectrally Selective Dithering for Distorted Speech Recognition”. English. In: *INTERSPEECH 2015*. Dresden, Germany, 2015. URL: http://www.isca-speech.org/archive/interspeech_2015/i15_2858.html

- A. Kolman P. Mizera P. Pollák and M. Ernestus. “Impact of Irregular Pronunciation on Phonetic Segmentation of Nijmegen Corpus of Casual Czech”. English. In: *Text, Speech, and Dialogue. 17th International Conference, TSD 2014*. Lecture Notes in Artificial Intelligence. Brno, CR, 2014, pp. 499–507. ISBN: 978-3-319-10815-5. DOI: 10.1007/978-3-319-10816-2_60. URL: http://link.springer.com/chapter/10.1007/978-3-319-10816-2_60

- M. Borský, P. Mizera, and P. Pollák. “Noise and Channel Normalized Cepstral Features for Far-Speech Recognition”. English. In: *Speech and Computer*. Lecture Notes in Artificial Intelligence. Pilsen, CR, 2013, pp. 241–248. ISBN: 978-3-319-01930-7. DOI: 10.1007/978-3-319-01931-4_32

- P. Mizera and P. Pollák. “Accuracy of HMM-Based Phonetic Segmentation Using Monophone or Triphone Acoustic Model”. English. In: *Applied Electronics - 2013 International Conference on Applied Electronics*. Pilsen, CZ, 2013, pp. 181–184. ISBN: 978-80-261-0166-6

Other publications

- M. Kosek and P. Mizera. “Implementation of Cepstral Voice Activity Detector”. English. In: *Proceedings of the International Student Scientific Conference Poster â 21/2017*. Praha, CZ, 2017, pp. 1–4. ISBN: 978-80-01-06153-4. URL: http://radio.feld.cvut.cz/conf/poster/proceedings/Poster_2017/Section_IC/IC_070_Kosek.pdf

- P. Mizera. “Rozpoznávání neformální řeči na bázi artikulačních příznaků”. Czech. In: *VI. Letní doktorandské dny 2016*. Praha, CZ, 2016. ISBN: 978-80-01-05959-3

- M. Borský, P. Pollák, and P. Mizera. “Vplyv ztrátovej kompresie v úlohe rozpoznávania spojitej reči.” Slovak. In: *V. Letní doktorandské dny 2015*. Praha, CZ, 2015, p. 63. ISBN: 978-80-01-05749-0. URL: http://sami.fel.cvut.cz/LDD15/Sbornik_LDD2015.pdf

- P. Mizera. “Implementace LVCSR s nástroji Kaldi v úloze rozpoznávání neformální řeči”. Czech. In: *V. Letní doktorandské dny 2015*. Praha, CZ, 2015, pp. 64–65. ISBN: 978-80-01-05749-0. URL: http://sami.fel.cvut.cz/LDD15/Sbornik_LDD2015.pdf

- A. Brich, J. Fiala, and P. Mizera. “Telephone Speech Recognition Using Time-Domain IIR Filter Bank in MFCC Computation”. English. In: *Proceedings of the 19th International Scientific Student Conferenece POSTER 2015*. Praha, CZ, 2015, pp. 1–4. ISBN: 978-80-01-05499-4

- J. Valíček and P. Mizera. “Language models for spontaneous speech recognition”. English. In: *Proceedings of the 19th International Scientific Student Conferenece POSTER 2015*. Praha, CZ, 2015, pp. 1–4. ISBN: 978-80-01-05499-4

- P. Mizera and P. Pollák. “Estimation of Articulatory Features for Czech Language”. English. In: *22nd Czech-German Workshop on Speech Communication. Book of Abstracts*. Praha, 2014, pp. 25–26

- P. Mizera. “Detekce a přínos artikulačních příznaků v úloze rozpoznávání spontánní řeči”. Czech. In: *IV. Letní doktorandské dny 2014*. Praha, CZ, 2014, pp. 81–84. ISBN: 978-80-01-05506-9

- P. Mizera. “The contribution of differential cepstral coefficients within AF estimation”. English. In: *POSTER 2014 - 18th International Student Conference on Electrical Engineering*. Prague, CZ, 2014, pp. 1–4. ISBN: 978-80-01-05499-4

- P. Mizera. “Zlepšení přesnosti fonetické segmentace na bázi HMM s akustickými modely trifónů”. Czech. In: *III. Letní doktorandské dny 2013*. Praha, CZ, 2013, pp. 92–97. ISBN: 978-80-01-05251-8. URL: `http://sami.fel.cvut.cz/LDD13/sbornik_LDD_2013.pdf`

- P. Mizera. “Mismatch Effect in Preemphasis Application within Speech Recognition Systems”. English. In: *POSTER 2013 - 17th International Student Conference on Electrical Engineering*. Prague, CZ, 2013, pp. 1–4. ISBN: 978-80-01-05242-6

- P. Mizera and P. Pollák. “Odhad základního tónu řeči s lokalizací hlasivkových pulsů a pitch-synchronní segmentace”. Czech. In: *20th Annual Conference Proceeding’s Technical Computing Bratislava 2012*. Praha, CZ, 2012, pp. 1–8. ISBN: 978-80-970519-4-5

Applied results

- P. Pollák, P. Mizera, and M. Borský. *Rekonstrukce řeči z reprezentace na bázi mel-kepstra II: optimalizace implementace*. Czech. Research Report FZ-2014-1. ZOOM International s.r.o., 2014, p. 4

- P. Pollák, M. Borský, and P. Mizera. *Tvorba akustických modelů na bázi HMM pro češtinu*. Czech. Research Report FZ-2013-1. ZOOM International s.r.o., 2013, p. 4

- P. Pollák, P. Mizera, and M. Borský. *Rekonstrukce řeči z reprezentace na bázi mel-kepstra I: základní implementace*. Czech. Research Report FZ-2013-2. ZOOM International s.r.o., 2013, p. 4

List of candidate’s work non-related to the thesis

Conferences

- A. Nautsch, J. Patino, A. Treiber, T. Stafylakis, P. Mizera, M. Todisco, T. Schneider, and N. Evans. “Privacy-Preserving Speaker Recognition with Cohort Score Normalisation”. English. In: *INTERSPEECH 2019*. Graz, AU, 2019. URL: `https://arxiv.org/abs/1907.03454`

- T. Stafylakis, J. Rohdin, O. Plchot, P. Mizera, and L. Burget. "Self-supervised speaker embeddings". English. In: *INTERSPEECH 2019*. Graz, AU, 2019. URL: https://arxiv.org/abs/1904.03486