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

Prosody Utilization in Continuous Speech Recognition

Doctoral Thesis

Ing. Jan Bartošek

Ph.D. programme: Electrical Engineering and Information Technology
Branch of study: Electrical Engineering Theory
Supervisor: Ing. Václav Hanžl, CSc.

Prague, August 2016

Thesis Supervisor:
Ing. Václav Hanžl, 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 2016

.....
Ing. Jan Bartošek

Abstract

This doctoral thesis covers the theme of prosody utilization in automatic recognition of continuous speech. Even though automatic speech recognition (ASR) systems have imoproved immensely over the last several decades, they still lack making use of one of the most important aspect of information using speech, which is a prosody. There have already been proofs from other languages about the favourableness of prosody usage in ASR and doctoral thesis tries to investigate the potential of Czech regarding prosody usage.

The research activities can be divided into three main areas: a) pitch detection algorithms (PDA) as needed prerequisite for prosodic feature extraction, b) Czech lexical stress system as potential clue from acoustic signal for word boundary detection (and its usage in ASR) and c) classification of sentence/phrase modality in Czech based purely on an acoustic signal.

Firstly, the field of pitch detection algorithms, a framework for their evaluation and comparison is presented. Several new evaluation criteria are proposed as an extension to existing ones together with metrics evaluation over four speech pitch reference databases. Besides pure comparison, few modifications of existing PDA methods are presented. Namely a transition probability function in PDA post-processing is investigated in terms of candidate distance measure and new temporal-forgetting principle for speech is brought in as extension of method by time domain.

Czech as a fixed-stress language with lexical stress on the first syllable is known to have a weak lexical stress acoustic correlation. Nevertheless, methods of how stressed syllables or stress-group boundaries can be detected from speech signal were investigated. A system with sophisticated feature extraction followed by statistical machine learning methods to model those phenomenon in Czech is presented. Detected stress-group boundaries can be (in most of cases) mapped to word boundaries which can be used for prosodic evaluation of ASR hypothesis. A metric for such prosodic score, which can be directly used in prosodic N-best evaluation or ASR error detection, is proposed. Also, ASR lattice rescoring algorithm for Czech is presented.

Czech phrase modality detection from acoustic signal is covered and together with existing phrase boundary detector can such system serve as an punctuation module for Czech dictation ASR system or in Czech dialogue system to support its natural language processing (NLP) part.

Keywords: Prosody, speech technology, ASR, F0, pitch, lexical stress, stress group, modality, melodeme, prosodic hypothesis scoring.

Abstrakt

Předkládaná dizertační práce se zabývá tematikou využití prozodické informace souvislé řeči v jejím automatickém rozpoznávání (ASR). Ačkoliv ASR systémy prošly dlouhou cestou ve svém vývoji, stále typicky postrádají využívání jednoho z nejdůležitějších informačních aspektů řeči, kterým je prozodie. Existující studie prokazují vhodnost využití prozodie v ASR pro jiné jazyky, v této práci je v tomto ohledu zkoumán potenciál češtiny.

Výzkumná část práce se zabývá třemi hlavními doménami: a) algoritmy pro odhad základní frekvence signálu jakožto nutné prerekvizity pro extrakci prozodických příznaků, b) systémem lexikálního přízvuku v češtině jakožto potenciálního zdroje informace úzce spjaté s hranicemi slov v řečovém proudu (a využití v ASR systémech) a c) klasifikace modality fráze v češtině výlučně za použití akustické informace.

V oblasti algoritmů pro detekci základního tónu (PDA) je nejprve představen framework pro jejich testování a porovnávání. Je navrženo několik nových hodnotících kritérií jako rozšíření stávajících spolu s jejich vyhodnocením přes čtyři různé F0 referenční řečové databáze. Kromě pouhého porovnání existujících PDA je předloženo i několik modifikací algoritmů samotných. Konkrétně je zkoumána pravděpodobnostní přechodové funkce v post-zpracování výstupů PDA ve smyslu míry hodnocení vzdálenosti jednotlivých kandidátů a navržen je nový princip "časového zapomínání", který rozšiřuje funkci o časový rozměr.

V češtině, jakožto v jazyce s pevným lexikálním přízvukem na první slabice, je tento přízvuk označován z hlediska akustických prostředků jako slabý. Nicméně, v práci jsou zkoumány metody detekce přízvučných slabik nebo hranic mluvních taktů z českého řečového signálu. Je představen systém se sofistikovanou extrakcí příznaků následován statistickými metodami strojového učení pro modelování tohoto fenoménu v češtině. Detekované hranice mluvních taktů se ve většině případů dají mapovat na hranice slov, které mohou být použity k prozodickému ohodnocení ASR hypotézy. Je navržena metrika takového prozodického skóre, která je přímo použitelná k prozodickému ohodnocení N-best výstupu rozpoznávače nebo pro detekci chybných ASR hypotéz. Nakonec je pro češtinu představen algoritmus pro přehodnocení rozpoznávací mřížky na základě prozodické informace.

V práci je také pokryta detekce modality českých frází je a spolu s již existujícími detektory hranic frází lze vytvořit interpunkční modul pro český diktovací ASR systém nebo jako nápomocný blok v českém dialogovém systému na úrovni sémantického zpracování a porozumnění přirozeného jazyka (NLP).

Keywords: Prozodie, řečové technologie, ASR, F0, výška tónu, mluvní takty, slovní přízvuk, modalita, melodém, prozodické ohodnocení hypotézy.

Acknowledgements

I would like thank to my family for supporting me from the very beginning of my university studies, my soulmate Jana (besides her great vegetarian lunches) for traveling with me through our lives and to all of my friends that were supporting me.

My thanks go to my supervisor Václav Hanžl for his support, countless meetings and brainstorming sessions we had, even after his departure from the academical sphere. I especially appreciate Unix/Linux shell and text-filter tricks he showed me occasionally on our way. I was able to adopt many of them and they helped me in writing endless amount of various scripts related to this work.

Special thanks belong to Helena Katzenschwanz (born Spilková) for her time she dedicated to our interesting phonetical discussions, for providing me with speech material from her diploma thesis and for always being a constructive critic of my occasional pure engineering views on speech.

Special thanks also goes to my colleague Petr Mizera for providing me with various force-alignments and for help related to Kaldi oriented tasks.

I would also like to thank Petr Pollák for his consistent practical writing/presenting tips and his time he dedicated to me.

There are surely many more people that helped or supported me during this long road of doctoral study. To these people, I apologize for not listing them here by name.

The whole dissertation thesis and all the related work tended to be based on the usage of open-source tools and toolkits, namely: Linux operating system (specifically Ubuntu and Debian distributions), BASH, awk (and all other Unix filters), perl, octave, Praat, R, Transcriber, WaveSurfer, gnuplot, Kaldi, HTK, Weka, ALSA, jack, sox, Mediainfo, Wine, Krusader, mc, Tex&LaTeX, Kile and many more. I would like to thank all the developers for creating all those wonderful pieces of software and for setting up the open-source community.

List of Tables

List of Figures

Contents

Chapter 1

Introduction

Speech has always been the primary way of human communication. For most of us, it is still also the fastest and the most natural way to communicate. As technology improves, humans are being replaced in various working tasks by machines and computers. It is, therefore, more and more common that people need to interact not only with each other but also with those programmed devices. For making this communication as easy and convenient as possible, there has been a great deal of effort in research of human-computer oriented speech technologies during past decades, and it still continues.

The automatic speech recognition (ASR) systems have two main domains of application. Firstly, they are the core of speech-to-text (STT) systems responsible for transcribing typically longer parts of human speech into texts in dictation systems or automatic subtitle generation systems (e.g. real-time live subtitling of television programs for impaired people). Secondly, they enable us to control electronic devices by speech in voice-control domain. Text-to-speech (TTS) systems oppositely transform written text into sound and thus allows computers to "speak". Both ASR and TTS systems are usually incorporated together as basic modules of super-ordinate dialogue systems, which connects them into the working aggregate while adding the logic of driving the machine-human based dialogue.

The crucial part of successful transcription of speech or communication between human and computer is the accuracy of ASR system. The main structure of ASR systems has stabilized over the years as methods that were most suitable to model individual aspects of human speech have been found and incorporated into the working whole. One of the standardized sections of overall ASR system flow is speech signal parametrization. Nowadays, almost all of used parametrizations try to disregard the differences between speakers and speaking styles. In this way, the following ASR acoustic models can be very general and whole speech recognition process is quite robust. Nevertheless, in real state-of-the-art ASR systems we still find various transcription errors. The causes of those ASR

errors vary from the insufficient amount of exact or similar examples in training data to speaker specific or non-standard realizations of training data. It is also shown that untypical prosodic properties of speech (especially intonation, loudness, speech rate), compared to typical training data, increases the chance of words being misrecognized. This explains the effort for disregarding of those prosodic properties (mainly fundamental frequency which contributes to perceived pitch) during the signal parametrization. On the other hand, it can be claimed that the prosody information is still useful cue for ASR systems - even for enhancing its output with punctuation marks or decreasing its word error rates (WER) and thus increasing its accuracy.

It is widely accepted that speech intonation (or melody) plays the most important part of speech prosody. To being able to use the intonation information in computer processing, an estimate of the fundamental frequency information from voice signal is needed to be done firstly. Although there exist wide range of methods for this task, their further research is still open because of significant demands on their high precision, robustness and low latency.

The thesis is organized as follows: Chapter 2 provides an overview of current state of the thesis problems followed by defined goals. In chapter 3 , a prosody in general with a focus on a Czech prosodic system is described. Chapter 4 is dedicated to pitch detection algorithms (PDAs), their evaluation and suggested modifications. Used prosodic material is presented and discussed in chapter 5. In chapter 6, problems of stress-group detection for Czech is described. Phrase modality task classification is investigated in chapter 7. Integration of examined prosodic aspects into ASR system is discussed in chapter 8. In the final chapter conclusions are summarized with a comparison of defined and realized goals.

Chapter 2

Current state

In this chapter the brief overview of problems is described with a current state of the research (where needed). Firstly, speech formation, hearing and its common acoustic descriptors are presented. Next, a concept of ASR systems is introduced. Later, the known ways of utilization of prosody in automatic speech recognition and speech related technologies are described. Finally, the thesis goals are presented.

2.1 Speech

Over the years, human speech had developed from more basic sounds which animals use to communicate between each other and became the main informative mean for humans. As the presented research is oriented on speech signal, its formation and perception is firstly shortly described. Further, speech signal acoustic descriptors are presented with the effort to distinguish between objective and subjective description.

2.1.1 Formation of speech signal

The creation of speech signal formation originates as exhaled air is pushed from lungs (due to muscles contraction after breathing in) upwards against the gravity into larynx. It passes through the vocal cords (vocal folds) placed inside the larynx where it might get periodic nature voiced sound if they are vibrating, or it just passes by as noise characterized sound without any fundamental frequency. After this, it pervades the system of head cavities. Their shape, which is adjustable by surrounding muscles, influences the final speech frequency content as signal is filtered by the resonant frequency of those cavities. This leads to emphasis of so-called 'formant' frequencies in the signal (usually denoted as F1, F2, F3 and F4).The first two formant frequencies play key role in vowel identity. The overall scheme of speech production system is depicted in the figure 2.1.

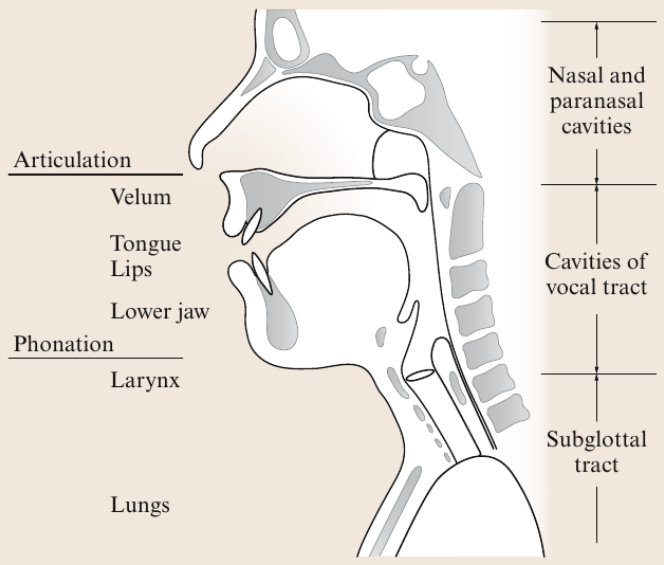


Figure 2.1: Sketch of speech production system (source [1])

2.1.2 Physiology of hearing

As the work presented in this thesis is very closely related to the human perception of acoustic (and particularly speech) signals, a brief overview of physiology of human hearing mechanism is described.

The sound comes initially into outer ear, then via middle ear and finally reaches inner ear where it is transformed into neural signals that are passed into brain and create the perception. Pinna is the only visible part of outer-ear and its shape influences mainly positional aspects of hearing for frequencies higher than 500 Hz. The outer ear continues with tube shaped ear canal (meatus) whose length corresponds to its first resonating mode at around 3 kHz. This resonance peak of height around 15 dB is important for intelligibility of speech signal. Ear drum (tympanum) divides outer ear from the middle one, where three ear bones hammer (malleus), anvil (incus) and stirrup (stapes) are located and where a center of effective acoustic conversion system from air environment into fluid environment (impedance adaptation) between outer and inner ear is. The middle-ear has also mechanism to protect our hearing for high intensity sounds if needed. The border between the middle and inner ear is created by oval and round windows. The main organ in the inner ear is spiral-shaped cavity cochlea with its tube divided into two floors by the divider with the sense organ of Corti. This organ contains hair cells that are able to pass the information to brain via ca. 30,000 threads of auditory nerve (using low-voltage electric impulses). The different frequencies excite different part of hair cells and brain obtains information about particular frequency content of the stimulus.

For a comprehensive description of the whole physiological process of hearing (and beyond) readers are recommended to follow [3].

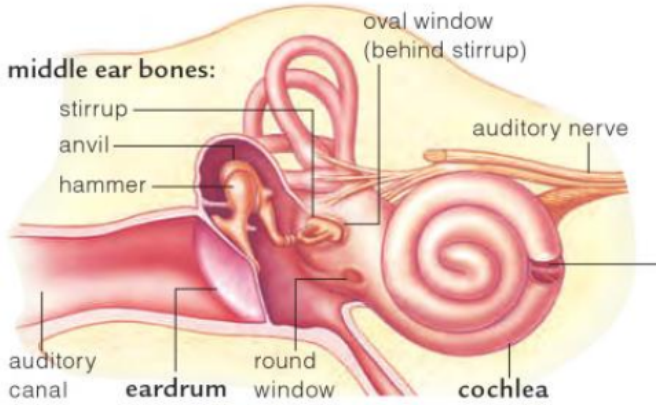


Figure 2.2: Ear scheme (pinna as the entry point of hearing is missing in the illustration and would be depicted at the very left side if present, source [2])

2.1.3 Properties of speech signal

Although there are many possible speech sounds which can be produced, the shape of the vocal tract and its mode of excitation change relatively slowly in time. Speech can thus be considered to be quasi-stationary over short periods of time (of the order of 20-25 ms) [4]. Voiced parts of speech signal can be defined from the physiological acoustic point of view as complex tones (periodical non-sinus signals composed by more frequency components). These basic properties denote the digital processing of speech signal, particularly its typical frame size and validity of usage of Fourier transform and other signal transforms.

2.1.4 Common speech signal descriptors

Commonly, three basic types of signal descriptors can be extracted from the signal. Pitch information describes the fundamental frequency for signals that contain it (e.g. voiced parts of speech). A strength of all signals can be described by its intensity. Spectral descriptors try to capture the frequency content of the signal and they usually pick just the most important events in signal spectrum and thus compress the full information. All of the listed types (or particular subtypes) find its usage as prosodic features.

Fundamental frequency

The physiological fundamental frequency F0 is the objective quantity that has its objective correlate in perceived voice pitch. It is obtainable directly as inverse of time distance T_0 (pitch period) between neighboring glottal-pulse pitch marks (e.g. result of laryngograph signal processing) or estimated as the output of pitch detection algorithm (PDA). The unit of F0 measure is Hertz [Hz] (number of periods within one second, $[s^{-1}]$).

Frequency measure in Hertz does not correspond to human perception of pitch which

follows Fechner-Weber's law. The law says, that the velocity of the sense is proportional to the logarithm of the stimulus (or alternatively, if stimulus increases geometrically, then the sense rises arithmetically). This is why conversion from absolute frequency units into relative musical units should be always performed in any task that tries to imitate or evaluate human perception.

$$ST_{diff} = 12log_2\frac{f_2}{f_1}\,[\mathrm{ST}] \tag{2.1}$$

Equation 2.1 defines a conversion from frequency difference in Hz units into semitone difference (musical scale) using 2-based logarithm of frequency ratios. As one can see, it is conversion from absolute to relative measure and a choice of the point f_1 of relativity is needed. To maintain defined "absoluteness" it is advantageous to relate the musical units to some common frequency unit. Common choices are 1 Hz or 100 Hz. The equation for conversion into semitones related to 100 Hz ($ST_{rel100Hz}$) is defined in 2.2.

$$ST_{rel100Hz} = 12log_2\frac{f_2}{100}\,[\mathrm{ST}] \tag{2.2}$$

Unit semitone [st,ST] corresponds to the musical unit of western music where interval of one octave (interval between musically "same" notes) is equivalently divided into 12 semitones. If finer measurement is needed, one semitone cent unit can be used (where one semitone consists of 100 cents).

According to several studies, it was proven that human hearing ends to have logarithmic nature at the frequency around 800 Hz and becomes "non-linear" compared to the logarithm. This is why several scales were suggested as even better approximation of fundamental frequency perception (mel, bark, erb scale). In this thesis the semitone scale is rather used because its straightforward conversion mechanism from frequency and due to the speech F0 ranges below the 800 Hz threshold.

The typical Czech men speech range is from 80 to 200 Hz (corresponding to cca. 16 semitones) with typical value around 120 Hz [5]. Women tend to have higher F0 range than men approximately by 9-10 semitones in average (which corresponds to the musical interval of major sixth or minor seventh), which roughly corresponds to the speech range of 150-350 Hz with typical value of 225 Hz. Children's voices are even higher with typical value of 300 Hz and voice range from 200 to 500 Hz. For comparison, singing range is wider than the speaking one. It is extended less to the lower frequencies (where speaking range already often uses the lower pitch limiting), but typically extends the upper boundary. Non-trained singers perform typically in the range of 1.5-2 octaves, while

trained singers can go up to 4 octave range (such training mainly improves the diaphragm support for achieving better exhalation control and pressure while keeping vocal chords relatively relaxed).

There are two fundamental frequency period stability descriptors usually measured only on long sustained vowels: jitter and shimmer. Jitter stands for frequency stability in terms of equality of periods duration and is computed as average absolute difference between consecutive periods, divided by the average period. Shimmer, on the other hand, presents measure for amplitude stability of F0 periods and is computed as average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude. Both measures have their empirically based thresholds and are mainly utilized in speech pathology research, where typically the third measure harmonics-to-noise ratio (HNR) characterizing voice hoarseness is also used (vowel identity dependent, too low values generally indicates a present hoarseness in voice).

Regarding the psycho-acoustics of pitch perception, just noticeable difference (JND) for pitch varies with frequency (the higher the frequency, the smaller the JND), but it is claimed to be about 1% for higher middle range of human voice (300 Hz). Thus, there are about 6 JNDs in one semitone around this frequency range (the next semitone is always about 5.95% higher than the preceding semitone). Several studies suggest greater JND for speech pitch perception as interval of a quarter step (half of semitone, 50 cents) [6], which is also base interval of micro-tonal music scale which contrasts conventional western-world semitone music composition and feeling. There seems to be general consensus about this lastly mentioned JND across phoneticians.

Noticeable speech pitch movement or glissando threshold (g_{thr}) is the minimum pitch change in semitones per second that can be distinguished from a tone of constant pitch. It was experimentally verified that it is dependent on the duration of the tonal speech sound T and the formula varies from $g_{thr} = 0.16/T^2$ to $g_{thr} = 0.32/T^2$ [7] saying that the longer the tonal speech sound the lower pitch change per unit time is needed to be perceived (decrease is non-linear with square of duration T). For the latter formula this means the g_{thr} being about a third of semitone per second for 1s long tonal speech sound and approximately 35 semitones (almost three octaves) per second for 100 ms long tonal speech sound.

Also, *differential* glissando threshold defining the minimum difference in slope necessary to distinguish between two successive glissandi (or change of pitch direction within a syllable) is defined and lies in between 12 and 40 semitones per second.

For the notion of the practical needed data bit-depth to encode typical speakers pitch values in his range, let's consider the JND of 20 cents (so there are 5 just noticeable differences in one semitone) and typical speakers speech pitch range of one octave (12 semitones). For that particular choice there are 60 different pitch levels (advantageously related for example to speakers baseline F0), so there is need for 6 bits (which gives $2^6 = 64$ different levels) to encode one particular pitch value knowing the speakers overall statistics.

Sound intensity and loudness

A perception of sound intensity (and also sound pitch) is driven again by Fechner-Weber's law. While intensity is objective and measurable description of sound, loudness is a psycho-physical sensation of perceived physical intensity by the human auditory perception (ear/brain) mechanism. There are objective measures for sound intensity description and also subjective quantities expressing its human perception.

Two measures are commonly used for objective sound level. The first one is coming from the definition of sound intensity as its true physical meaning (being emitted power into unit area $[W/m^2]$), while the second comes from sound pressure as slight atmospheric air pressure modulations (which is the principle of signal capturing by deviation of microphone diaphragm and thus is directly related to captured voltage and captured audio signal amplitude). Both of them use 10-based logarithm of ratio as core for the transformation from original unit into the 'level' measure. Sound intensity level (SIL) is defined as:

$$L_I = 10log\frac{I}{I_0}\,[\mathrm{dB}] \tag{2.3}$$

where I_0 corresponds to the reference value of $10^{-12}W/m^2$.

Sound pressure level (SPL) is on the other hand defined as:

$$L_p = 20log\frac{p}{p_0}\,[\mathrm{dB}] \tag{2.4}$$

with p_0 reference value corresponding to the threshold of hearing (for continuous tone with frequency between 2 and 4kHz) at $2.10^{-5}Pa$. The intensity of the sound wave is proportional to the square of the sound pressure ($I \propto p^2$), which explains the multiplying constant difference after logarithm application in equations (2.3) and (2.4).

Generally, it can be claimed that SPL and SIL levels match to each other. Strictly speaking, the equality between sound intensity level (L_I) and sound pressure level (L_p) is limited by the condition of exact acoustic impedance of air $z_0 = 400kg/m^2s$. Nevertheless,

for ranges of barometric pressure between 990-1040 hPa and temperature between -30 and +40 °C their difference does not exceed the value of 0.2 dB [8].

The strength hidden in usage of decibel units might not be clear as usual linear difference in basic units is replaced by ratio of original unit for the decibels, which rises new "kind" of arithmetics and understanding. This, for example, means that 20 dB increase can be caused either by ten-time increase of sound pressure or by hundred-time increase of its intensity. Doubling the source of signal (e.g. two loudspeakers playing instead of just one) generates a decibel level increase by only 3 dB. Increasing the number of players from 1 to 10 brings an increase of 10 dB, which is considered in psycho-acoustics as the moment when perceived loudness is approximately doubled. If the amplitude (voltage) of the audio signal is doubled an increase of relative level by 6 dB is obtained.

JND of intensity varies across frequencies (in the frequency range of 500Hz-4kHz we are able to distinguish up to between 260 loudness levels), but currently the general consensus is considered to be the intensity level change of 1 dB.

Subjective perception of loudness is largely dependent on sound frequency as the sounds of the same intensity cause various loudness senses for various frequencies. This is why various experiments were performed to obtain the curves of equivalent perceived loudness (known as Fletcher-Munson's curves [9]) by comparing the examined sine wave intensities and frequencies with those at 1kHz. As defined like this, perceived loudness level in phons [Ph] at 1kHz directly corresponds to objective sound pressure level (SPL).

It is worth noting, that those curves are valid for sine signals, but the reality is even more complicated as perceived loudness is also affected by other parameters than sound pressure and frequency including bandwidth, spectral composition, information content, time structure, and the duration of sound signal exposure.

Nevertheless, none of the earlier presented level measurments unfortunately match the sense of perceived loudness. This is why new measure *loudness* was created and anchored to loudness level of 40 Ph. Following equation defines loudness N in [sone] units as derived from loudness level L [Ph]:

$$N = 2^{\frac{(L-40)}{10}} \text{ [sone]} \tag{2.5}$$

The advantage of using loudness measure lies in the fact that naturally expected principle additivity of stimuli and the perception is valid here (compared to any dB and Ph measures). Particularly, an increase of loudness level by 10 Ph causes perceived loudness in sones to be doubled and decrease by 10 Ph leads to halving of loudness value. The practical impact of conversion mechanism is demonstrated in the table 2.1

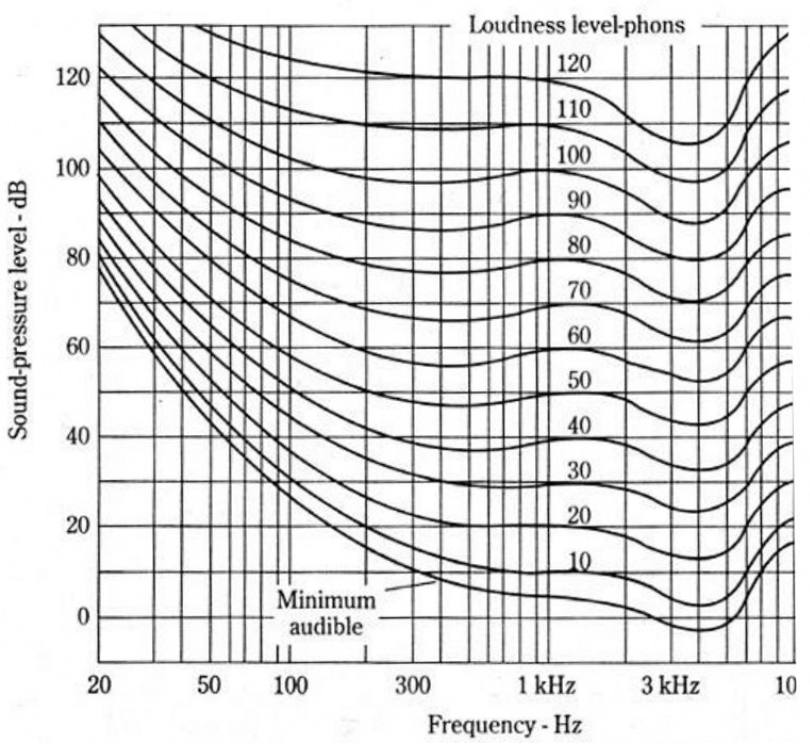


Figure 2.3: Fletcher-Munson's curves of equivalent perceived loudness [9]

L [Ph]	20	30	40	50	60	70	80
N [sone]	0.25	0.5	1	2	4	8	16

Table 2.1: Conversion table between subjective Phon units of perceived loudness level L and perceived loudness N in sone units (according to the equation 2.5). Loudness level range of normal talking from 1 m distance is from 40 to 60 Phons.

There is very little doubt about the advantages of using pitch represented in musical units (semitones, cents) in any human-perception oriented tasks because differences in all those measurments almost perfectly match the perception of the examined acoustic quality. On the other hand, in terms of sound intensity and its perceived loudness relation, it cannot be definitely claimed, that perceived loudness defined as in sone units is the primary choice as an energetic feature difference descriptor of signal in any machine learning tasks (even if they can be obtained objectively).

In the world of time discrete digital signals the energy is defined as

$$E_s = \frac{1}{N} \sum_{n=0}^{N-1} s[n]^2 \tag{2.6}$$

Equation 2.6 is a base for obtaining an energetic signal envelope, that has various utilization in speech processing. In prosody it is particularly connected with detection of lexical stress accent realized in many languages by energy increase during stressed

syllables. Logarithmic scale is also commonly used in connection with energy of speech signal.

Spectral characteristics

The most suitable parametrizations in terms of machine audio signal processing seem to lie in utilization of the information contained in the signal spectrum. The spectrum is obtained using Fourier transform, in digital signal processing mainly by its faster implementation known as Fast Fourier transform (FFT). In both cases the result consists of two parts, magnitude spectrum and phase spectrum, both can be expressed as single complex number. In terms of speech processing, it is the magnitude spectrum which is mainly utilized including the information about which particular frequencies are contained in the signal.

Since the magnitude spectrum contains the most complete information about signal, there is tendency to compress the whole spectrum image into a low dimensional space (especially during ASR signal parametrization in frontend). Moreover, full magnitude spectrum often contains too detailed information and some smoothing techniques available via its parametrizations are suitable. In the ASR field there are two most used speech signal parametrizations: PLP and MFCC. Both still well enough characterize the original spectrum and are presented shortly.

- Spectral slope
  Spectral slope is the most simple spectral characteristics consisting of the only value. It is defined as ratio or difference of two separate frequency band energies. It is often used to describe vowel character in terms of the effort of its production (higher frequencies are increased more than lower frequencies with more effort). More details on spectral slope within speech prosody can be found in section 3.1.4.

- Formants
  As the basic concept of formant frequencies was already described in section 2.1.1, they correspond to resonance frequencies of vocal cavities and first two of them are typical descriptors of vowels. In the magnitude spectrum of signal they are often represented by clear peaks surrounded by deeper valleys. From four to five formant frequencies are being tracked usually.

- Perceptual linear prediction (PLP)
  PLP was firstly presented by [10] and extends conventional linear predictive (LP) analysis. It uses three concepts from the psychophysics of hearing to derive an estimate of the auditory spectrum: the critical-band spectral resolution, the equal-loudness curve and the intensity-loudness power law. The auditory spectrum is then

approximated by an auto-regressive all-pole model. PLP analysis is computationally efficient and yields a low-dimensional representation of speech.

- Mel-frequency cepstral coefficients (MFCC)
  The most common final domain of extracted features is Mel-frequency cepstral domain and features are then denoted as Mel-Frequency Cepstral Coefficients (MFCCs). The steps needed to obtain these coefficients c_n from original signal $s[n]$ can be seen in the Fig.2.4. In short, mel-frequency cepstrum represents a short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel-scale of frequency. Mel-scale is motivated by perceptual masking phenomena, when non-linear bank of filters approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. Computational details together with other used speech parametrization techniques can be found e.g. in [5].

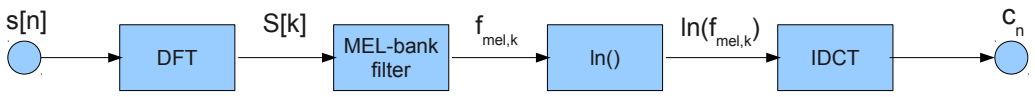


Figure 2.4: Block scheme of MFCC coefficients calculation [5]

2.2 ASR systems

The role of automatic speech recognition (ASR) system is to convert human spoken speech into recognized textual form (often suitable for further machine processing). In the statistical approach to the problem, the task of speech recognition can be mathematically described by equation 2.7.

$$P(\hat{W}|\mathbf{O}) = \arg\max_{W} P(W|\mathbf{O}) = \arg\max_{W} \frac{P(W)P(\mathbf{O}|W)}{P(\mathbf{O})}, \tag{2.7}$$

where $\hat{W}$ is sought word sequence, $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_N\}$ is the sequence of observations derived from acoustic signal (features) and $W = \{w_1, w_2, ..., w_M\}$ is a word sequence. $P(W|\mathbf{O})$ denotes a probability of word sequence W given the acoustic observations $\mathbf{O}$ which can be after application of Bayes rule rewritten as a fraction, where $P(W)$ is the probability of word sequence W regardless the acoustic information (coming purely from language model), $P(\mathbf{O}|W)$ is the probability of acoustic observations given the sequence of words W (coming from acoustic model) and a priori probability of acoustic observation $P(\mathbf{O})$ being constant can be omitted due to argmax function.

2.2.1 Methods and blocks used in ASR

There are various blocks and related machine learning methods even in the standard ASR system. The most important of them are briefly described with an attempt to capture the underlying basic ideas in an understandable way. The basic scheme of simplified ASR system is depicted in the figure 2.5.

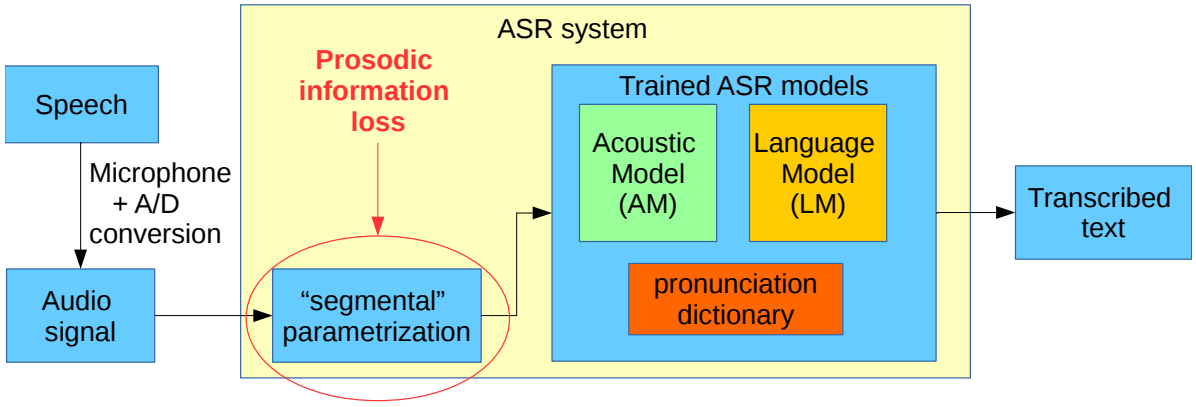


Figure 2.5: Simplified diagram of general ASR system with targeted prosody information loss during audio signal parametrization

Frontend

The "frontend" is the first block in the ASR system and is responsible for acoustic pre-processing (like optional de-noising, etc.) and following feature extraction/signal parametrization (mandatory). A noise reduction block added before feature extraction block leads to better robustness of the system against environmental issues (common method is spectral subtracting of actual noise profile obtained from silence passages). Reverberation added to the speech signal caused by non-reflective surfaces nearby the speaker in combination with non-close-talk microphone (distant channel) presents serious difficulty for the ASR (compared to music, where reverb is very often added to make close-miked singing of instrument sound more pleasant to human ear). Methods for de-reverberation are being actively developed to overcome this issue.

Typically, spectral information of signal is extracted here using MFCC or PLP features (described in section 2.1.4), which are usually computed each 10 ms using the frame duration of 20-25 ms. Rate of frame overlapping influences time precision of processing (length of frame on the other hand influences the frequency precision). In the beginning is the signal frame weighted by windowing function (Hamming, Hann, etc.) to avoid occurrence of spectral leakage phenomena after conversion to spectral domain. The temporal scope of 20-25 ms is suitable to capture spectral content of actual single phone (or a transition between two phones at most) as this is exactly the duration for which speech

signal is considered to be stationary. Thus, the frontend parametrization is often called as segmental and it is exactly this part of ASR system, where the very most (except energetic feature) of prosodic information gets lost. This loss can be considered as wanted and needed in order to be able to reach the highest speaker independence of consequently trained acoustic model. Common dimension of a single frame feature vector, containing except raw values also first (velocity, delta) and second order (acceleration, double-delta) derivatives, is 40 (3x13 cepstral parameters + 1 energy parameter). LDA transformation is often used to project data from high dimensional feature space (capturing also the context information) into reasonable dimensionality (~40D).

Acoustic model

Acoustic phoneme level matching classifier is the next level of parametrized speech signal processing. The acoustic model can be trained from acoustic speech data with labeled phonetic transcription to recognize how each phoneme really sounds, which in many cases depends on the surrounding phonemes (there is contextual dependence). That is why triplets of phonemes called "triphones" are often used as basic unit for the acoustic model. Because the phoneme realizations in feature vector space are not exactly the same even for the same speaker saying twice the same utterance, the phonemes are statistically modeled. Two classification method based on statistical models are currently used, older GMM and newer and more powerful DNN.

- GMM (Gaussian Mixture Model) is a parametric probability density function. In ASR it models probability in the N-dimensional space of features, where total probability is derived as mixture of single probabilities in each dimension modeled by Gaussian curves, each with own mean and variance. The output probability density function $b_i(\vec{o})$ (Eq. 2.8) assigns a probability of acoustic agreement of evidence (incoming feature vector) with GMM model over all N dimensions.

$$b(\vec{o}) = g.exp(-\frac{1}{2}\sum_{k=1}^{n}\frac{(o_k - \mu_k)^2}{var_k}), \tag{2.8}$$

where

$$g = \frac{1}{\sqrt{2\pi^n}\prod_{k=1}^{n} var_k}). \tag{2.9}$$

In the equation 2.8 ($\vec{o}$) denotes incoming n-dimensional feature vector, μ is mean of Gaussian curve and var its variance in given dimension. Logarithm of the whole

function $b_i(\vec{o})$ is often computed to simplify following computation (converts probabilities multiplication to addition).

- DNN (Deep Neural Networks) represent recent boom in all machine learning tasks. The main advantage of any modern neural network lies in its non-linear units that are (after training) able to model non-linear functions, which do occur in real world. Term "deep" comes from utilization of many hidden layers (typically more than 2), while classical neural networks tend to have only one hidden layer and are thus marked as "shallow". The deepness has been a problem for classical error-correction back-propagation learning/training algorithms based on stochastic gradient descent (SGD). The deep structure did not allow the derivation function to propagate through all the layers in a controlled way. In 2010 Hessian-free (HF) second-order optimization approach for training of deep neural networks was presented [11], which outperformed unsupervised pre-training presented in 2006 [12] as the first method to overcome the back-propagation problems in learning of deep structures.

The part of ASR system which gives result of acoustic similarity of unknown input audio data to known phones in context is called 'labeler'. For GMM based system, the hierarchical labelers are often used for real-time computation speed-up with clustering of phone Gaussian prototypes (first level of hierarchy is roughly sampled and every deeper level of hierarchy consists of more final set of phone prototypes).

The amount of training data differ according to their availability and used type of acoustic classifier layer. For GMM models it is suitable to train on hundreds of hours, while best results for deeper neural architectures are achieved on rather 1000+ hours of speech.

Time alignment using Hidden Markov Model

Hidden Markov Model (HMM) is present in the ASR system to cover the temporal relations on the segmental level (between individual phones in context). Being able to assign the phoneme to incoming feature vector is the first condition for successful phoneme alignment process. The second one reclines in the ability of modeling the variable phoneme (or in general every speech unit) length in real speech. For this purpose HMM finite state machine framework is usually applied. HMM is statistical tool for stochastic process modeling and can be viewed as a general extension to Dynamic Time Warping (DTW) approach also used for dealing with time issues of alignment. The Markov model has non-emitting initial and final states and an indefinite number of emitting states between

them. HMM of contextual-dependent phoneme usually consists of three states and can correspond to diagram in the Fig.2.6.

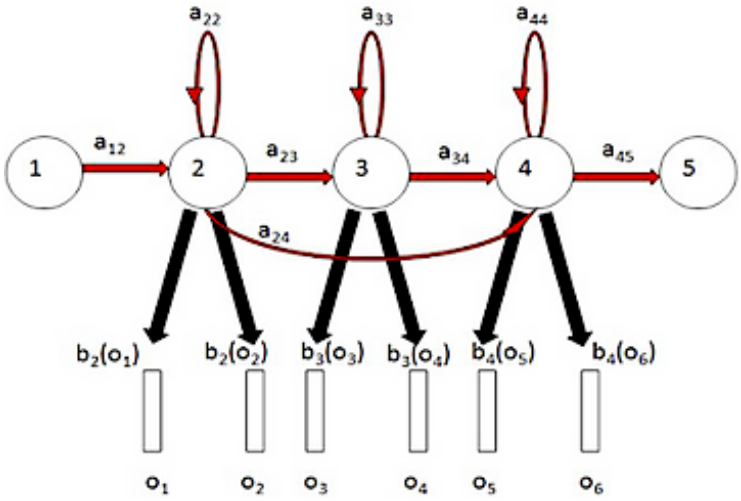


Figure 2.6: Illustration of Hidden Markov Model with three emitting states

Function $b_i(\vec{o})$ (equation 2.8) can be also viewed as the probability that in state i the feature vector $\vec{o}$ will be emitted. States of the model are traversed forward in time (only left to right transitions are permitted) and frames of feature vectors o_t are emitted. The probability of moving between one state i and the next state j, is given by probability a_{ij}. The initial and final non-emitting states are used for connecting of various HMM together and thus being able to model greater units like syllables, words or even the whole sentences. In a HMM, the state sequence, the transition probabilities a_{ij}, and the output probabilities $b_i(\vec{o})$ are unknown and must be educed from the training data.

GMM-HMM architecture of ASR system is usually adjusted by two global parameters: number of HMM states and total number of Gaussians that can model those states in the feature space. During the acoustic model training from transcript labeled acoustic data, all the parameters of acoustic model and also parameters of HMM are iteratively improved using information of segment boundaries from audio data which get more precise with every training iteration.

Speaker adaptation techniques

Among speaker acoustic adaptation techniques, most belong to fMLLR (Feature space Maximum Likelihood Linear Regression) and VTLN (Vocal Tract Length Normalization).

Although the MLLR transform was originally proposed as a model-space transformation for GMM individual speaker acoustic model [13], MLLR is nowadays frequently interpreted as an inverse feature-space transformation and marked by term constrained MLLR or fMLLR. The original model space MLLR is defined a speaker adaptation tech-

nique in which the means of Gaussians in an ASR system acoustic model are adapted so as to maximize the likelihood of the adaptation data for a particular speaker. On the other hand, in fMLLR we are trying to find the best adaptation (typically represented by a matrix) of speaker's features to best match the trained model.

VTLN presents a quite comfortable way of adapting the ASR system to the user. The idea comes from different vocal tract length between man and women (by 10-15% in average) which influences also different resonant frequencies of the "tubes". Those are projected into speech signal at formants at different average frequencies. The VTLN divides speakers into several (typically 15-20) categories and for each category (index) specific pre-computed transformation (exactly linear warping of the frequency axis) of features exists leading into unified feature-warped space in which the acoustic model is trained by standard methods. VTLN can be seen as an extension of gender specific models while dividing speakers into more than two categories. During the run-time, the unknown speaker is "initialized" with medium warping factor which converges into the real one as more speakers data are available for statistics. Frequency scale warping is often integrated into the filter-bank used to obtain MFCC features.

Training a discriminative model

Training a discriminative model presents alternative to classical generative models. As often abbreviated into term "discriminative training", the correct term for description of usage of discriminative methods should be "training of discriminative model" as it is the model whose properties changed, not the training methods. The main idea behind training of discriminative model is to choose the model parameters that improve the task and not those, that fit the training set best. Compared to a conventional training technique using maximum likelihood estimation (MLE), which attempts to maximize the training data observations likelihood, new objective functions as maximum mutual information (MMI), minimum word error (MWE) or minimum phone error (MPE) are widely used in training of discriminative model.

Language model

A lexical side of the speech recognition task corresponds to the allowed sequences of words on the recognizer output. They can be defined explicitly by form of a grammar, or on the basis of statistics and probability by language models (LM).Classically, N-gram models (typically up to 3-gram or 4-gram) capturing probabilities of occurrence not only for single token (word), but also of the tokens (words) in the context up to N tokens, are

used. Training of language model occurs only on text corpus with segmented sentences against a dictionary of allowed words (words not contained in the dictionary are typically unified into a single token "unknown"). To keep the language model in reasonable data size, the occurrence threshold filter for words is often applied and various smoothing techniques are applied to avoid zero-frequency problem and to balance the weights of words with high occurrence versus those with low occurrence in the corpus. Perplexity is the property of LM that is usually used for its evaluation. Perplexity can be viewed as a measure of predictability of the language. The lower the perplexity of LM, the better the results can be expected.

Language models (LM) can be either general or domain specific (e.g. banking sector, personal assistant, etc.). Also, regardless the domain, it is known that read speech (broadcast news) is different from spontaneous speech (dialogues, meetings) and this fact should be also considered in data corpus for LM training with respects to the planned application.

Decoding network

Final decoding (the real usage of trained ASR system on unknown speech signal, which is transformed into hypothesis on the output) with trained acoustic and language models is realized using either static or dynamic decoding network. Networks are constructed from the given trained language model, acoustic model and pronunciation dictionary. The decoding graph itself has multiple levels with word graph (coming from language model) on the top, phone graph (expanded from word graph using pronunciation dictionary) in the middle and state graph on the lowest level (expanded from the phone graph using HMM defined topology). In the run-time, the only input to the decoding network is audio signal processed by frontend. The task of ASR system in run-time can be then understood as finding the most optimal path through the graph. Decoding performance is always trade-off between accuracy and speed (usually tweakable via few parameters) but it has been shown that for GMM/HMM systems there is a threshold from which minimal gain is obtained even with much slower decoding speeds (non-linear function).

The result of decoding is usually internally stored as hypothesis lattice consisting of all hypothesis that "survived" during decoding process. This lattice can be exported for further re-scoring (typically with a language model with higher order of N-grams) or is just used to output demanded certain number N of best hypothesis contained in the lattice (N-best). The scale factor, which is a weight between the language model and acoustic model scores used to compute final hypothesis score from the lattice, is very important ASR setting as it largely influences the final accuracy. Thus, it is one of the hyper-parameters that are searched for its optimal value when combination of specific

acoustic and language model is going to be deployed, so the whole system can perform at its best.

ASR in 2016

Typical ASR systems in 2016 can be described as DNN based hybrid architectures (recurrent neural networks with Long-short term memory, convolutional neural networks) for capturing the acoustic space while still wrapped by HMM together with higher N-gram language models (up to 8-grams). DNN architectures allowed to significantly lower the WER on various benchmark tasks. For example, actual best results on English Switchboard database (one of the most favorite benchmarks of conversational telephone speech) reached a word error rate of 6.8% (while WER of 40% was achievable with best systems in 1995). It is estimated that human average WER on this task is around 4% [14].

The attention has been also attracted to the low-resource training of ASR systems, having available only tens of hours of speech, but using advanced and state-of-the-art modeling techniques. Also, conversion of trained ASR systems into different languages is very interesting and closely watched topic.

2.2.2 Error rate analysis of current ASR systems

Although thehe is a long history of ASR systems research and development, they still produce significant amount of errors depending on the task complexity, environment, etc.

To the best of our knowledge, the most thorough typology and examination of ASR errors can be found in [15]. While the primary aim of the study was to clarify common observation of remarkably varying word error rates among individual speakers, the clear causes of this phenomenon were not revealed. Individual word error rate (IWER) was defined as needed alternative to common WER to find the individual problematic words using two distinct ASR systems for English. The claims from other already existing studies were that the words that are not frequent are more likely to be misrecognized, fast speech (but occasionally also very slow speech) negatively influences the error rate. More errors can be seen on shorter words and more troubles have ASR systems with male speakers then with female speakers. Performance of the ASR system is also influenced by phonetically similar words, for which frequency-weighted neighborhood density coming from psycholinguistic field [16] is their reasonable joint predictor. The results supported all of those claims, except the last one. Authors in [15] argued that in the word context available in the ASR system, most of phonetically neighboring words are not problem for the recognition and present rather term 'doubly confusable pairs' which causes the

possible errors (homophones or neighbors with with similar context predictability).

From the perspective of presented thesis, prosody related features (next to the disfluency, categorical, probability and pronunciation based features) were also examined in [15] as individual category of error rates. The unit for prosody feature extraction was a whole word segment and various statistics covering pitch, intensity, speech rate and duration were followed. It was found that examined prosodic features were strongly predictive of error rates. Decreased duration was associated with increased IWER, high values of mean pitch across the words (relative to gender average) too. It would be nice to see logarithmic semitones values for relation to gender mean normalized features, because there are present examined words with mean F0 higher by more than 150 Hz from gender mean, which might indicate musical distance of 14 semitones if F0 mean was 120 Hz for men and 9 semitones difference if the mean women F0 was 220 Hz (particular gender means are not quoted). Words with smaller ranges of pitch or intensity were most likely to be misrecognized, as were the words with high minimum intensity feature. Jitter and intensity mean and maximum show instead pattern of higher error rates at extreme values than at its average values. To summarize the prosodic feature examination, extreme prosodic values led to more ASR errors. Since current speaker adaptive training techniques (MLLR and VTLN) are focused on cepstral content of speaker differences on segmental/phone level, they are not able to cover prosody differences. Thus, suggested is speaker-adaptive training to deal with prosodic variation with possible requirement for explicit representation of prosodic aspects. This suggestion has been already partially covered by [17] where speed of audio signal is changed for acoustic training in various levels. Soon, there can be expected to appear similar works on data modification in other prosodic manners to generate prosodically rich material with even extreme realizations and make ASR systems more robust.

Unfortunately, neither any words insertion/deletion issues are described in [15], nor any other prosody-based suggestions for ASR error improvements are considered.

Prosody oriented research of errors produced by French ASR was done in [18]. By the mapping of syntactic lexical units into stress-groups (in the text, they are marked as prosodic words, PW) according to know rules for French. In French, the lexical stress belongs to the very last syllable in the stress-group and should be contrasted by the pitch increase of more than 2 semitones. According to this rule, they automatically marked H* pitch accents from the acoustic data. They found, that only 43% of the expected stress-groups were actually detected in the signal using this procedure. One of explanation was, that many syntactic units were too short to consistently map to own prosodic words. Lastly, they provide qualitative study of two ASR errors that can be discovered by

using the extracted prosodic information about stress groups (ASR hypothesis do violate the suggested stress-group boundaries obtained from acoustic signal).

Work [19] brings (to the best of our knowledge) the first study of ASR errors for real Czech ASR system aimed for massive transcription of Czech broadcast archive. The study is focused more into alignment issues between reference and hypothesis on the sub-word level covering even portions of the words and consequent substitutions, deletions and insertions are examined in more detail. Firstly, main sources of ASR errors (consisting of imperfect lexicon, inadequate AM and LM, speaker's style and noisy environment) are presented. To my knowledge, none of those problems can be solved by using prosodic information. The work continues in introducing a scheme that offers more detailed insight into ASR analysis based on 763 hours of manually labeled versus automatically transcribed data. Their error location method based on might be helpful also in other Slavic languages, mostly because of detecting the errors coming from the wrong pronunciation lexicon. From the perspective of this thesis, the most important seems to me their example of ASR error with a hypothesis "... sedákem balili můžeš na místě" againts the reference "... se také bavili muži z námestí". It is interesting in terms of Czech enclitic 'se' in the reference transcript, which would by any chance create a beginning of any true stress-group, while in the wrong hypothesis the word 'sedákem' will very probably create its own stress-group with first stressed/accented syllable. The prosodic acoustic analysis can in this case very likely reveal potential error in the ASR hypothesis. On the other hand, prosody is not limitless and again, due to Czech clitics behavior it would be very likely not able to distinguish between collocation "na místě" in the end of hypothesis and correct reference "náměstí" as monosyllabic preposition 'na' being a proclitic will very likely become a first syllable of joint stress group " 'namístě ", which will be acoustically not distinguishable in terms of stress from the reference. More information on clitics and stress-groups can be found in the sections 3.4.2 and 6.2.

The conclusion from this section is, that in real state-of-the-art ASR systems there are still real transcriptions errors that can be solved by utilizing the prosody information.

2.3 Prosody utilization in speech-technology

This section covers broader context of current prosody usage within various speech technologies (excluding the ASR systems, which are discussed in more detail separately in the section 2.4). There are various tasks that directly correspond to prosody of speech and direct utilization in speech technologies. Those topics will be mentioned only briefly as

they lie on the border of the scope of presented thesis.

Nevertheless, I will try to pick some of them and describe their current state of the art as I think that it might be the kind of information, which some readers of this thesis could be interested in. The overview marks out the importance of prosody as a cross-discipline field of research with various and often overlapping applications.

2.3.1 TTS systems

Text-to-Speech (TTS) systems have been probably on of the first speech technologies, where prosody usage was inevitable from the very beginnings due to the fact that direct output of TTS system is perceived by the human users who have comparable perception of real human voice. Initial TTS systems were based on vocal tract models and true voice synthesis, but although a huge effort they never reached bigger popularity due to unnaturalness of resulting speech. Current approaches rather lie in concatenative synthesis, where real chunks of recorded speech from the TTS database are being used in various contexts. The true magic is in appropriate selection of those chunks from the database and modification of their prosodic values to match the neighboring context (especially PSOLA pitch shifting), so that the result still sounds natural.
TTS systems have to basically deal with 3 main prosodic 'levels' in given (non-tonal) language: Naturalness of stress-group prosody (various acoustic qualities might be involved) and natural overall sentence prosody (here mainly intonation). Those two phenomena are composed together using the principle of superposition, while natural sentence-level prosody is achieved by higher acoustic contrasts then prosody on the stress-groups level. In addition, natural/wanted emotional state of speaker (also carried by various acoustic qualities) is nowadays a wanted commodity of TTS systems and can be achieved both by database coverage or speech signal acoustic qualities modifications.

The summarizing guide to prosody for real Czech TTS system was carried by [20]. The presented underlying technology was implemented into Czech version of Epos TTS. Sound qualities used for presented prosodic modification of original synthesized speech are only F0 contour and duration, while aimly not changing the intensity contour on this "general prosody" level by any means (which is in accordance with their previous experiments). It presents TTS aimed prosody modeling on the level of 'stress units' while firstly presenting set of rules for their formation from the text). The "Initial", "Middle", "Final" and "FinalFinal" stress units are distinguished according to their position in the cluase-unit, while each of them having specific "models". Lastly, the chaining of stress units into intonation phrases is covered Against classical approaches, they bring new view on the problems by not distinguishing special sound characteristics for the first syllable in

the stress-unit, but their approach is rather based on patterns (which corresponds more with observed reality for Czech stress).

The importance of prosody in TTS proves the Speech Synthesis Markup Language (SSML). It is a XML-based mark-up language for description of TTS prompts standardized by W3C organization. It contains special prosody attributes influencing: pitch (baseline [Hz] or relative change), pitch contour (sequence of target values at specified time positions), pitch range (range [Hz] or relative change), speaking rate (relative change or wanted absolute speed), duration (absolute time) and volume (absolute or relative change) The most of current open-source and commercial TTS system do support the SSML standard.

To judge the overall quality of TTS system, an evaluation metrics exists for that purpose. The most common one is subjective Mean-Opinion-Score (MOS) estimation (1-5 scale, where 1 is BAD and 5 is excellent) Common MOS of recorded real human utterances are usually between 4.6 and 4.7. Current best TTS systems achieve MOS score around value 4.0. Interspeech 2014 TTS evaluation contest was criticized by [21] for too low number of listeners particularly in MOS tests and suggest the minimum value as 30 so the results are statistically significant.

The main focus in TTS is currently still on naturalness of speech achieved mainly by pitch (and other acoustic qualities) contour modeling. The recent studies use advanced models as Long Short-Term Memory (LSTM) Bi-Directional Deep Recurrent Neural Networks [22]. Current effort in TTS systems development also lies in the ability of expressiveness or emotional feeling of speech.

2.3.2 Detection of emotional state of speaker

Humans sometimes behave emotionally which also reflects their nature of speech. An interesting topic of research is to find out, which emotional category corresponds to particular speaker's utterances (emotions usually rise from dialogue with other person or machine).

The task is not only pure theoretically-oriented area of research, but has direct practical real utilization, even in business sphere. One of the actually solved and crucial use-case is in contact centers. During the call with live agents (operators) the system might detect if the customer is getting angry on the operator and if this is the case, he might suggest to the operator

HMM-based English system for emotion classification was presented in [23]. It is based on contours of time-domain glottal source model and filter parameters, which were found to be significant for naturalness of emotional speech synthesis. System was evaluated on Emotional Prosody Speech and Transcripts LDC corpus that contains 'acted' emotional speech collected from seven professional actors with the speech comprising preselected, semantically neutral phrases. Best results were achieved by a combination of all available feature contours and a confusion matrix can be found in the table 2.2. Although the reported overall accuracy 62.6% may not seem very high, human listeners achieved average accuracy of 63.6% on the same LDC corpus (11 listeners were involved in the listening test).

	Neutral (%)	Anger (%)	Sad (%)	Happy (%)	Bored (%)
Neutral	**55.3**	0	23.4	4.3	17.0
Anger	0	**81.7**	0	16.9	1.4
Sad	3.3	0	**57.4**	14.8	24.6
Happy	0	23.3	5.5	**60.3**	11.0
Bored	3.9	0	32.5	7.8	**55.8**

Table 2.2: Confusion matrix for the HMM-based English emotions recognition system using all parameter contours, (overall accuracy 62.6%, according to [23]).

For Czech, the state-of-the-art publications is [24] to the best of my knowledge. Next to the classical prosodic features (F0, intensity, duration) plays quite a key role also spectral content of the signal. Gaussian Mixture Models (GMM) were used as a modeling framework. Achieved results in terms of mean error rates for classification of 4 basic emotions can be found in the table 2.3.

Emotion:	Neutral	Joy	Sadness	Anger
Male	25.64	7.14	7.69	43.65
Female	7.69	39.39	3.45	33.33
Total	16.67	23.28	5.57	38.99

Table 2.3: Mean error rate results for Czech emotion classification (according to [24]).

2.3.3 Prosody-based speaker recognition/verification

Speaker recognition/verification has variety of applications. Except its security applications it features also forensic utilization and ASR utilization, where the speaker identity information is beneficial in terms of usage of particular (already trained/estimated) models or characteristics in speaker adaptation techniques to increase the recognition accuracy.

Comprehensive overview and novel approaches regarding the speaker verification even on prosody level is given in [25]. On the other hand, global phonetic view on the problem of speaker identification brings [26] with summary of recent experiments carried out for Czech. Considering the prosody markers, F0 does not provide currently any relevant and reliable criteria except statistical indicators (mainly baseline F0 and even it does not serve as identity confirmation, but rather as identity exclusion). The application and parametrization of melodic contours seems to be quite a difficult task. There is possibly some potential for speaker identification/verification in individual prosodic models covering individual acoustic realizations of speaker's stress-groups. On the other hand, temporal indicators as articulation rate and rhythmical descriptors (especially "%V" standing for percentage ratio of vocalic intervals in the speech) seem to be very robust for the given task and are resistant even against aimed fraud voice masking. Descriptor %V allowed to significantly distinguish 39% of compared speaker pairs (out of 561 pairs in total).

2.3.4 Language recognition

In the field of language recognition, the main event regarding this topic was definitely "2009 NIST Language Recognition" contest. The task was to classify 23 of languages, Czech was unfortunately not included. Surprisingly, the prosody and phonotactics seem to be more a supplementary source of information for the classification task, rather than the main cues as expected. Nevertheless, few works are oriented anyway in prosody usage to fulfill the given assignment [27], [28]. The state-of-the-art approaches are based rather on advanced acoustic modeling techniques using recurrent neural networks with LSTM units [29].

2.4 Prosody utilization in ASR systems

Prosody usage in speech recognition system is not a new idea and various attempts were realized in history. In this section, a presentation of existing ways of its usage in ASR systems is given.

In the figure 2.7 the overall diagram is presented of general ASR system enhanced by prosody as it is seen from the perspective of this thesis. The following existing approaches described subsequently might or might not correspond to the given diagram. Also, topics in many of cited works do overlap and thus the work may have belong to another category of prosody utilization in ASR too.

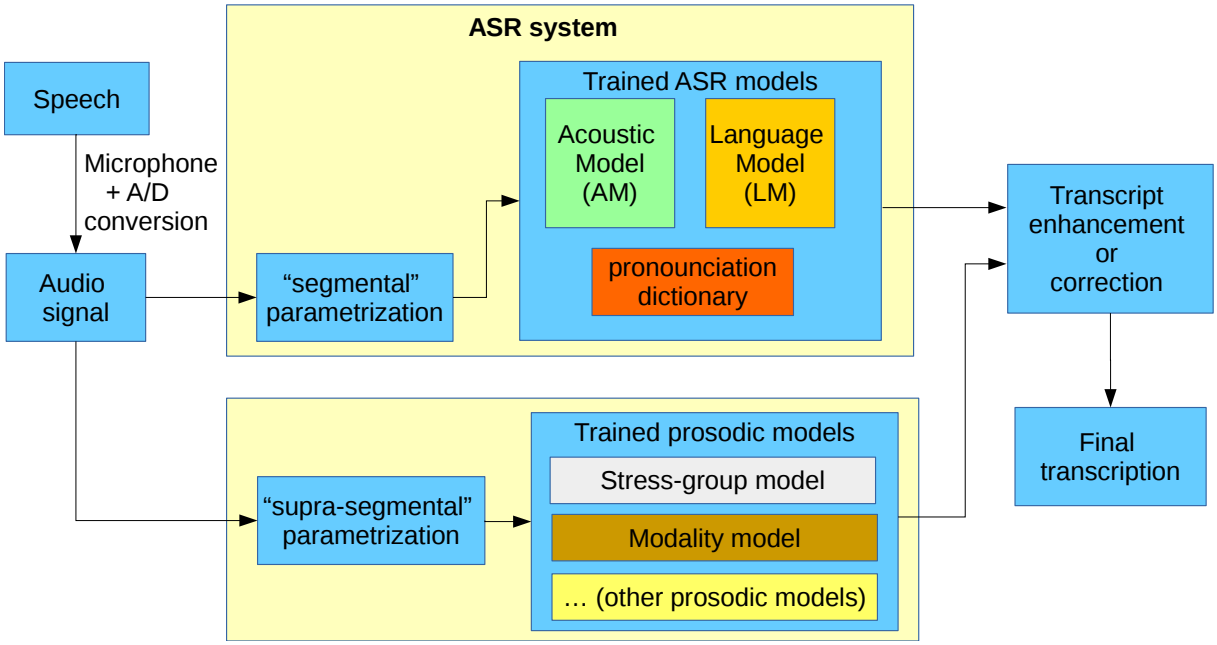


Figure 2.7: Simplified diagram of prosody assisted ASR

2.4.1 Hypothesis rescoring

The first successful attempts of using prosody in ASR systems were probably done in 1993, when N-best rescoring algorithm based on prosodic features was presented [30] and used in travel service application. They used decision-tree classifiers for prosodic breaks discovering and classifying the prominence of each syllable. A speech synthesis tool was used to generate references for all the N-best hypothesis, that were compared with detected output break and prominence sequence. Average rank of hypothesis chosen by their classifiers was improved significantly. Many of applications using prosody in ASR were implemented within VerbMobil project [31] and dealt with sentence or clause boundary detection. [32] used MLP neural networks to recognize prosodic phrase boundary. They also tried to incorporate phrase boundary events into language model instead of tracking the prosodic information from speech signal itself. Both methods reduced WER of the ASR. Syntactic units boundary detection and hypothesis rescoring is addressed in [33]. Next to the MLP approach they investigated special poly-gram language model containing some prosody events. This Study [34] examined Japanese prosody rich system and used it for word-boundary detection, but their method is strictly bound to Japanese word-level intonation system. They used HMM models to model intonation contours.

2.4.2 Punctuation detection

There are several studies dealing with punctuation detection. The first of these studies used only lexical information by building 3-gram language model [35] (and recently [36] with dynamic conditional random fields approach), others also utilized acoustic information [37], when acoustic baseforms for silence and breath were created and punctuation marks were then considered to be regular words and added to the dictionary. Another Study [38] concluded that pitch change and pause duration is highly correlated with position of punctuation marks and that F0 is canonical for questions. CART-style decision trees for prosodic features modeling were used. In [39] a detection of three basic punctuation marks was studied with combination of lexical and prosodic information. Punctuation was generated simultaneously with ASR output while the ASR hypothesis was re-scored based on prosodic features. Ends of words are considered as the best punctuation candidates. For this reason, all the prosodic features were computed near the word ends and in two time windows of length 200 ms before and after this point. The prosody model alone gives better results than the lexical one alone, but the best results were achieved by their combination. Authors also mentioned complementarity of prosodic and lexical information for automatic punctuation task. Combination of prosodic and lexical features also appeared in [40] where punctuation process was seen as word based tagging task. Pitch features were extracted from a regression line over whole preceding word. Authors also mentioned evaluation metric issues and except for Precision and Recall (P&R, F-measure), they also used Slot Error Ratio (SER) as well. Language model in combination with prosody model reduces P&R and SER, especially with the pause model for full-stop detection.The maximum entropy model was presented for punctuation task as a natural way for incorporating both lexical and prosodic features in [41], but only pause duration was used as prosodic features. Lexical-based models performed much better then pause-based models which is in contrast to the other former studies. Work [42] presents an approach for punctuation based only on prosody when utilizing only two most important prosodic attributes: F0 and energy. Method for interpolating and decomposing the fundamental frequency is suggested and detectors underlying Gaussian distribution classifiers were trained and tested. In [43] the idea continues and it is claimed that interrogative sentences can be recognized by F0 (intonation) only and about 70% of declarative sentences can be recognized by F0 and energy. A closely related task to the automatic punctuation is sentence boundary detection which is discussed in [44], where a pause duration model outperforms language model alone. Again, the best results are achieved by combining them.

2.4.3 Unsupervised adaptation of categorical prosody models

The study [45] deals with unsupervised adaptation of categorical prosody models for prosody labeling and speech recognition. The prosodic subsystem of ASR is presented as a special separate couple of acoustic and language models. These categorical prosody models are decoupled from baseline ASR so that they can be implemented and trained independently and their final advantage is baseline ASR lattice enrichment. The major shortcoming of categorical prosody model is the lack of large enough training corpora with symbolic prosody labels. This can lead into poor estimate of model parameters and into high Out of Vocabulary (OOV) errors in the case of prosodic language model. The authors have solved this issue by presenting the method of unsupervised model adaptation technique for both prosody models using large prosody unlabeled database. The adaptation method was then evaluated by pitch accent classification. For the training of the seed models, the Boston University Radio News Corpus (BU-RNC) [46], consisting of 3 hours of read news annotatied in ToBI style (pitch-accents and boundary tones), was used.

The adaptation technique for prosodic language model (PLM) is based on confidence weights generation by seed model for adaptation data from the adaptation dataset and following the estimate of new language model that is merged with original one. This resulted in a 13% improvement in relative reduction of binary pitch accent classification. Authors have also presented an algorithm for adaptation of prosodic acoustic model (PAM) in a way very similar to the PLM adaptation. They obtained full-covariance PAM that decreased the pitch accent classification error by 4.3% relatively. Both of the proposed adaptation schemes do not employ any discriminative adaptation techniques. The adapted prosody models were finally adapted into ASR system. With incorporated prosody models, the relative WER error decrease was 3.1% with lattice enrichment.

2.4.4 Prosody-assisted Mandarin ASR

In the field of tonal languages, where the pitch information mainly phonological role (it distinguishes between word meanings), the most comprehensive work on a prosody-assisted Mandarin ASR was presented in [47]. In the new probabilistic framework for Mandarin Chinese speech recognition incorporates sophisticated hierarchical prosody model into the conventional hidden Markov model based recognizer. Four-layer prosody hierarchy is used to generate 12 models are generated from unlabeled speech database by the joint prosody labeling and modeling algorithm, which they presented earlier. Word lattice rescoring based on prosodic model can lead to better recognized word strings. Besides those seg-

mentation errors, the system was able to correct also tone recognition errors. They experimentally verified a significant improvement on Mandarin TCC300 database consisting long paragrahpic utterances from 24.4% WER to 20.7% WER (15.2% relatively).

2.4.5 Prosody utilization in Hungarian ASR

An article [48] is a complex work on utilization of prosody in ASR systems. It is oriented primarily on Hungarian language, but also involves Finnish and German in particular experiments. The aim of the work was to show that the examination of prosody characteristic is advantageous in speech processing and their involvement to the automatic speech recognition is useful. Being still the state of the art paper in 2016 for fixed-stress languages, we are going to describe used methodology and experiments in more detail.

Modeling presented in the study is based only on acoustic prosodic features, but finally influences also lexical level of the process. This Study presents multi-level speech recognition system, because prosody can contribute to syntactic and semantic level processing in ASR systems. At the syntactic level, authors claim that prosodic segmentation allows word-boundary detection and following rescoring of word hypothesis graph based on prosodic information. But firstly, word-stress unit alignment is needed. Word-stress units consist of single word or word group corresponding to speech syntactic unit. This kind of improvement led to higher recognition accuracy in real medical system. On the other hand, at the semantic level of processing, prosody information can be used for sentence boundary detection and sentence modality recognition. In this case an alignment of modality type prosodic models (which model different modalities) is needed for speech. Having the time-alignment of the clauses and sentence boundaries, we can place the punctuation mark corresponding to aligned sentence modality model.

The prosodic information extraction is based on continuous prosodic contours classification on both levels of processing. For the statistical modeling of prosody at both levels, common and "practice-proven" HMM framework was used, usually used for speech unit alignment in speech recognition task. Adoption of this technique in supra-segmental domain allows the direct transfer of prosody information into common ASR system. The work also suggests prosodic acoustic feature extraction and pre-processing needed for both tasks. In Hungarian, only the fundamental frequency and energy are considered as main supra-segmental features. Fundamental frequency is the most important cue for the stress in Hungarian language, whereas duration-like features were not found reliable for fixed-stress Hungarian in previous research [49] and thus they were discarded from the work. Fundamental frequency was obtained by short time Average Magnitude Difference Function and together with energy were measured in 10 ms intervals on frames with 25 ms length. F0 values were processed by octave filter and then linearly interpolated into loga-

rithmic domain, but leaving pauses longer than 250 ms. In the last step of pre-processing, the mean filter was used on both pitch and energy contours, but with different processing length. In syntactic module where it is needed to preserve stress and filter only micro-intonation, relatively short window was used (5-point mean filter). For this case, first and second deltas of both F0 and energy were computed and final 6-dimensional feature vector is obtained. In the semantic module there is a need to apply longer window to eliminate word-stress and keep only modality-related part of intonation pattern. This is why 5-50 point mean filters were used. Also, feature vector consisted of 14 number in total, representing 3 various intervals for computing first and second derivatives of both basic prosodic features. In both cases these supra-segmental features obviously do not correspond with features used in classical ASR.

Word boundary detection at the syntactic level is based on and idea, that in fixed-stressed languages the stress occurs always at the well-defined part of the word or within the whole group of the words.This stress can be aligned with use of extracted supra-segmental features to the input speech. When the stressed units are carefully labeled and then identified so that their boundaries correspond to word-boundaries, the alignment gives the word boundaries. Used trained HMMs are then able to model different stress realizations using extracted F0 and energy. Given syntactic level module has then similar structure to the traditional ASR with the only main difference that modeled entities are word-stressed units and not speech units (phonemes, syllables, etc.) as usual. This also means, that special grammar (having the same role as language model in common ASR) is also needed to express allowed sequences of word-stress unit models.

The example of common Hungarian utterance and its prosodic feature contours are shown in the Fig. 2.8. Fundamental frequency (F0) and energy are measured in the middle of vowels, durational feature represents the vowel (nucleus) duration. It can be seen that peaks in F0 and energy clearly correspond to stressed first syllables of the words. But it is also true that not all the words within the sentence are stressed. According to the Fig. 2.8 one can also see that stress in Hungarian is expressed more significantly by F0 and energy. Durational features were found to be unreliable for Hungarian.

Special supra-segmental annotated databases are needed to train the prosodic models for both the tasks. For the first task, Hungarian BABEL (1600 sentences of 32 speakers) was segmented on the basis of prosodic features for word-stress units by an expert. Annotation technique differed from methods that label high or low accents (such as the ToBI standard [50]) as the labeling used was simpler and thus less expensive, while still satisfied the needed task. Word-stressed units were segmented from stress to stress or from stress to phrase ending and can thus correspond not only to single word. Finally, five different types of word-stress units according to fundamental frequency contour were found and de-

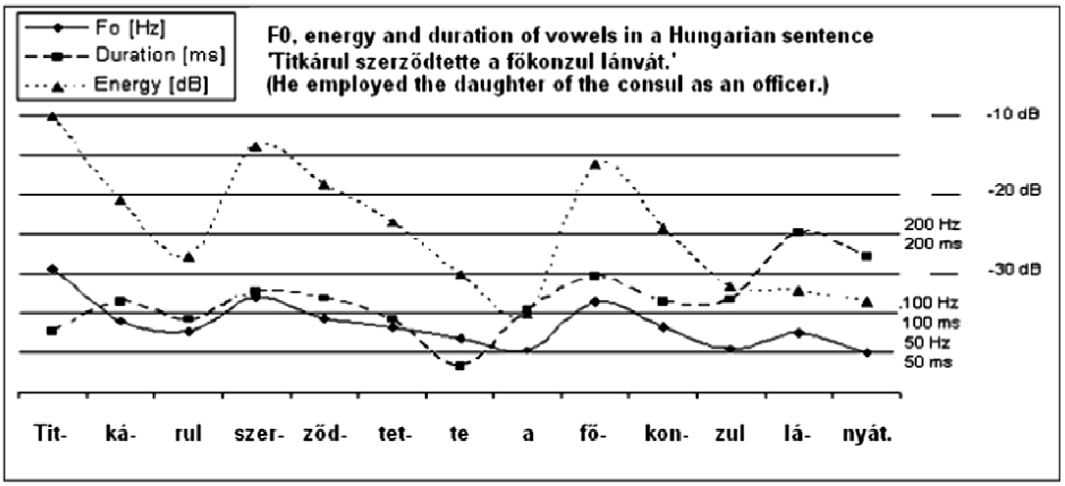


Figure 2.8: Hungarian sentence with corresponding F0, energy and vowel duration contour (source [48]).

cided to be modeled: descending (DE), falling (FA), ascending/rising (RI), rise-fall (RF) and floating-like (FL). Also, pauses and silence were modeled as special class too. Authors refer that in labeled material 80% of word-stress unit patterns were descending or falling, 10% was silence and the rest was rising, floating and rise-fall together. Thus, six different HMMs were constructed at the syntactic level. Authors also presented heuristic grammar describing allowed word-stress unit sequences, but mention only three classes (RF, RE, DI) out of five in it. For classes FA and FL it is probably difficult to educe any rules. Simply said, the presented grammar defines that speech is as sequence of word-stress units, which is a free combination of RF, DE and RI elements. Pause can be present after DE unit and is required after RI unit (denoting the end of prosodic phrase). As we are dealing with word boundary detection task, it is not so important the stress unit to be correctly assigned to certain class, but rather correctly aligned in time. Authors found optimal count of prosodic HMM model states (11) and number of Gaussian components sufficient to approximate their probability density functions (from 2 to 4). Correctness of prosodic segmentation is interpreted as the ratio of correctly detected boundaries and all detected boundaries in the validation set (more than half of all the labeled data). The tolerance range was set to 100 ms. The overall best results were 77% with models trained on only F0 and energy of 4 Hungarian speakers.

The final integration of syntactic module into ASR is advantageous especially for search space reduction for agglutinative languages (Hungarian and Finnish), where nouns might have more than 1000 different forms in combination with word relative free order in the sentence. These are serious problems leading to very large dictionaries and high complexity of statistical language models (in comparison to English).

Detected boundaries were also converted into probability density function $L_B(t)$ ac-

cording to its accurate position in comparison with reference. Word-boundary detector as the syntactic level sub-module works independently on the main ASR up to final N-best lattice rescoring. This lattice is created in the last step of common ASR module and is passed to the syntactic sub-module. The principal of the rescoring is to augment the scores of the candidates, whose boundaries match the detected word-boundaries. On the other hand, it punishes the hypothesis that contain detected boundaries within themselves. The rescored lattice is used as weighted grammar for the recognition in the final ASR stage. Whole system was evaluated on the Hungarian medical real-time application and absolute correctness of ASR system rose from 76% to 78.9%.

Unlike in stress-unit alignment, the result of classification of the sentence modality recognition module is important. Sentence modality HMMs map the prosody of speech represented by its F0 and energy to modalities. The semantic module realization is very similar to the syntactic one, but it needs rougher time resolution in comparison to syntactic module as mentioned before. Hidden Markov models for modality were trained on previously labeled databases. For Hungarian, sentence modality read speech database was constructed with six basic modalities. Sentences were split into clauses, where complex sentences have non-terminal and terminal clauses. The terminal clause was marked with proper sentence modality symbol, the non-terminal clauses were classified as separate group. An overview of used clauses and its corresponding intonation contours is in the Tab. 2.4. For each modality type a single HMM was trained plus one for silence/pause.

Clause modality	Intonation contour	Label
Declarative and declarative terminal	descending	D
Non-terminal	floating or floating-slow rising	NT
Wh-questions terminal	fall-descending	Q
Yes/No-questions terminal	rise-fall	YN
Imperative and exclamation	rise-descending	IE
Optative	ascending-descending	O

Table 2.4: Modalities of clauses and its corresponding intonation used for Hungarian database labeling (according to [48]).

Heuristic modality grammar containing simple rules was used again. Each modality is allowed to occur and can be preceded by optional pause/silence and no-terminal clauses. It was found that for modality recognition task, the span of mean filter did not essentially influence the results in the range of 100-500 ms. Again, the 11-state HMM gave the best results. The best average recognition of modality was 59% for Hungarian, while correctness of sentences and clause-boundary detection was 94%.

Presented syntactic level processing method is language specific, but is is claimed to

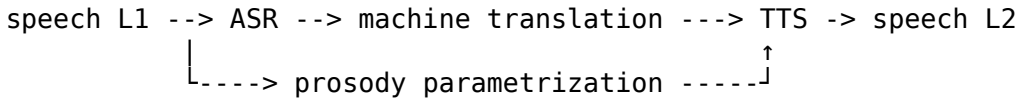
work on the whole group of the fixed-stress languages with stress on the first syllable (where Czech belongs). On the other hand, the models at the semantic level used for clause detection and sentence modality recognition should be language independent (the only condition is that language express modality by intonation) and probably there could be possible cross-lingual use when models trained on one language data are then evaluated on another language with reasonable performance. From the point of view of presented thesis, study [48] brings valuable framework which is worth to experiment with regarding the Czech language.

2.5 Prosody utilization beyond ASR systems

This section brings short overview of prosody utilization which is being currently worked on already or is expected to be investigated in the near future.

2.5.1 Speech-to-Speech translation systems

The aim of such systems is conversion of speech in original language (L1) into speech in target language (L2) with an effort to keep as much prosodic information as possible [51]. The whole process can be schematized:

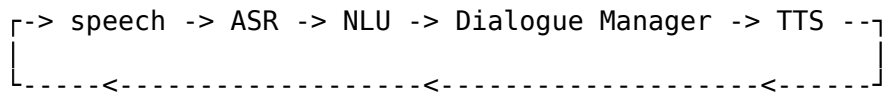


The underlying technology is definitively language dependent and the key factor is mapping of prosodic utterance structure from one language into another, even with different order of words in the sentence, etc. This kind of systems is surely a future of any teleconference communication services allowing a spoken conversation between people that might have otherwise no chance to understand each other. The real-time factor of conversion will certainly play a key role for communication fluency connected with cognitive load of such dialogue, too.

2.5.2 Dialogue systems and personal assistants

Those are systems with the aim on longer-term single-user usage. Those systems are most complete and pending to replace real human counterpart in communication. State-of-the art natural language understanding (NLU) systems used for entity extraction and overall information retrieval process from speech are followed by dialogue managers that are on the other hand responsible for choosing the right action given the historical context of

current communication and actual information obtained from the user. The whole process can be again schematized:



There are couple of places where prosody takes its part. It occurs on the system input (where there has been already attempts to its utilization [52]), and both on the output in synthesized TTS prompts (here it has been used from the very beginnings of TTS systems).

With the long term usage of that single-user, it will enable to create his adapted prosody models (e.g. personal stress-group and phrase intonation models), which will improve not only the ASR accuracy but also semantic extraction in natural language understanding process.

Also, there will be attempt for TTS prosody adaptation according to the obtained user's prosody models. This will enable mimicking for example a dialect of the user. It is known that participants of dialogue tend to adapt the acoustic/prosodic way of their speech to each other [53] and if this is not happening, the participants get the inner feeling of displeasing communication. It can be perceived as courtesy social act. Such TTS fine-tuning will lead to improved user's experience (UX) of system usage as the overall communication with artificial system will be inherently more pleasant for the user (subliminal stimuli).

2.6 Goals of the thesis

In stress-languages, speech stress helps native speakers understand each other, especially if the language is spoken quickly. On the other hand stress information is very rarely used in current automatic speech recognition systems. That is why presented work is going to be focused on the problems of stress detection and its utilization in Czech ASR. Also, tempting curiosity how much of the sentence or phrase categorical modality information is encoded in pure acoustic signal in Czech will be addressed.

To summarize, two main prosodic-oriented tasks/utilization suitable for enhancing the automated speech recognition systems are investigated: segmentation of Czech speech into stress group and sentence modality detection. Additional attention will be paid to pitch detection algorithms being a key methods for prosody information extraction. Therefore, the goals of the thesis can be defined as the following list covering three main research areas:

2.6.1 Pitch Detection Algorithms

- study existing pitch detection algorithms (PDAs) and create their overview
- examine the existing PDA evaluation criteria and suggest their suitable modifications that might improve the insight into performance of pitch trackers
- implement an evaluation framework over various available speech pitch reference databases and compare the algorithms
- think about PDA modifications for speech or music domain, try to implement them

2.6.2 Czech stress-group system

- study the known facts about Czech lexical stress and stress-group system
- study the Czech clitics rules, develop lexical tool for automatic sentence division into stress groups for Czech
- prepare prosodically enriched Czech database in terms of lexical stress suitable for machine learning processing
- consider suitable acoustic feature extraction and their normalizations for the task of Czech stress-group segmentation

- develop a classifier-based system for detection of Czech stress groups from speech signal that will use only acoustic information (no features or information from lexical-based modules except for vowel identities)
- propose an objective measure for evaluating ASR hypothesis in terms of its prosodic probability, which will allow N-best assessment
- experimentally verify the usability of suggested measure
- propose a scheme for Czech ASR lattice rescoring using prosodic information about utterance stress-group segmentation

2.6.3 Czech phrase modality

- study the known facts about Czech system of phrase modality
- prepare prosodically enriched Czech database in terms of phrase modality suitable for machine learning processing
- consider suitable acoustic feature extraction and their normalizations for the task of Czech phrase modality classification
- explore a classifier-based solution for Czech phrase modality detection using pure acoustic information (no features or information from lexical-based modules)

2.6.4 Notes and work demarcation

The work will not cover in detail the task of sentence/phrase boundary detection as this was already deeply studied in [54]–[56] for Czech and later in [57] for Hungarian as another fixed-stress language. In [58], performance for Czech was compared to English. The side result of presented work might be an extension of phonetic knowledge of above mentioned tasks in Czech, as those are so far described as "complex" and challenging.

Chapter 3

Prosody

When people convey information to each other by speech, it can be claimed, that prosody is one of the main cues. Prosody is often marked as suprasegmental characteristics of speech including speech intonation, rhythm, stress. Emotional state of the speaker can also be recognized because of global perception of individual prosodic qualities. One of possible definitions is also that prosody is information gleaned from the timing and melody of speech. As speech consists on its lowest level of individual segments (typically phones or phonemes), term suprasegmental denotes that prosodic events occur in the scope of multiple speech segments.

Prosody is a multi-disciplinary field of research. It forms the core of various research questions in a wide variety of fields, including linguistics, speech technology, psychology, and computer science [59].

The main fact that explains the attention of prosody as a separate chapter in this thesis is its possible usability in ASR systems as it presents additional acoustic information that was so far not much utilized. Although the prosody can be seen as subcategory of phonetic research, in this chapter will be also presented more general terms and facts related to phonology and possibly to phonotactics.

3.1 Prosody related acoustic features

Prosodic characteristics are achieved by modulation of various acoustic features that are perceived by listeners. It is suitable to follow the objective versus subjective terminology of those common features. Prosodic perceived correlate of fundamental frequency (F0) is called pitch (also intonation or melody), objective measure duration becomes subjective length, perceived intensity is denoted as loudness and spectral structure is subjectively referred to as a timbre.

In this thesis the usage of correct terminology is tried to be fulfilled, but for example the term "pitch", is used in technically oriented works it is still a synonym for fundamental frequency F0.

This is why term "pitch detection algorithm" (PDA, chapter 4) is used quite often although it does not correspond fully with the subjective perception. Rather, to be consistent with existing studies to some extent, in this work term *fundamental frequency* (F0) is used for measuring objective glottal frequency in Hertz units and the term *pitch* is used after logarithmic conversion into semitone based musical scale, which correspond much more with human perception mechanism. Nevertheless, it needs to be emphasized that even term pitch as used in this thesis still corresponds to objective measure rather than to true subjective perception. Its exact conversion into perceived intonation is still an open topic of research influenced by many psychoacoustic and context factors (author's personal experience) which are beyond the scope of this work, but known and used approximations are discussed further in sections 3.6 and 6.3.

Also to note, the term "intonation" is not being used uniformly as it has its wider and narrower meaning. In its wider sense it denotes a complex of speech melody, rhythm and stress, while in the narrower (and to our knowledge more commonly used) sense, it covers speech melody only. In this work it is used in its narrower sense as the descriptor for speech melody only.

The last general note belongs to just noticeable differences (JNDs), which were already covered for basic sound descriptors in the section 2.1.4 in terms of static acoustic signals. It is actually known that the well-established experimental methods developed in psychoacoustics do not actually apply in the field of intonation [60]. Study [61] investigated JNDs for Czech in terms of evaluation in global and local scope. It suggests that there is an incapacity of describing our own prosodic percepts when listening to continuous speech in Czech (global scope), but it does not mean that any perception occurred – prosodic parameters are integrated into patterns and embedded into linguistic units and the original prosodic details are being perceived unconsciously. In the local scope when only two neighboring syllables were directly compared without further context, no improvement was observed for evaluation of duration. The intensity evaluation suggest JND to be 4 dB for perceptual increase. Considering the F0 in Hertz, the asymmetric behavior was observed with JND of 8% for fall and 4% for rise. To summarize, the JNDs established for stationary signals cannot be directly applied in real speech prosody.

3.1.1 Fundamental frequency

Modulations of pitch, being perceived fundamental frequency (F0, section 3.6), create speech melody. If the pitch has phonological functions, it means it can distinguish between

word meanings. In the tonal language Mandarin Chinese, there are claimed to be 9 different tones which denote up to 9 different meanings of the same phoneme sequence differentiated only by its pitch in the non-tonal language understanding.

Even in non-tonal languages, speech melody carries an important part of speech information. There are two interfering phenomena regarding the melody; the first is its function on the word (stress-group) level and the second on the phrase or sentence level.

There are several phonation types voice from the fundamental frequency and glottal cycle point of view:

- normal voice:
  Chest voice register, vocal cords are in their natural mode (modal voiced phonation, contact between the vocal folds occurs).

- creaky voice, vocal fry:
  F0 below normal voice with very low frequency, vocal folds are strongly adducted and of weak longitudinal tension. Creaky phonation is characteristically associated with aperiodic glottal pulses. The degree of aperiodicity in the glottal source can be quantified by measuring the "jitter", the variation in the duration of successive fundamental cycles. Jitter values are higher during creaky phonation than other phonation types. Also, in creaky voice vowel the amplitude of the second harmonic is greater than that of the fundamental, which enables it to be differentiated by spectral tilt [62].

- falsetto voice:
  F0 is noticeably higher than in modal voice. The vocal folds are stretched longitudinally, thus becoming relatively thin. Consequently, the vibrating mass is smaller and the generated tone higher. The glottis often remains slightly open, resulting in low sub-glottal pressure (due to constant glottal leakage) and the generation of the audible friction noise component.

- breathy voice:
  Often marked as compound phonation type (voiceless+modal). Vocal fold vibration is inefficient and, because of the incomplete closure of the glottis, a constant glottal leakage occurs which causes the production of audible friction noise.

Local deviations of pitch up to 3 semitones are referred as micro-intonation [63]. Those changes occur unconsciously and are an articulation property of phonemes combinations. The physical trend here is that there is rise in F0 when the vowel follows an unvoiced explosive, and there is F0 decrease when it follows a voiced explosive). On the other hand

there exist also phenomenon of inherent or intrinsic F0, where the average F0 does differ for different vowels in the same sense and is dependent on the openness of particular vowel (the difference is up to 6% in frequency in French according to [63]).

3.1.2 Intensity

The acoustic intensity with derived sound intensity level and sound pressure level have been already covered by section 2.1.4 in physical and psychoacoustic terms. In term of analog or digital audio signals, it is the signal amplitude that corresponds to the intensity. From prosodic perspective, intensity of signal is perceived as loudness and plays various roles. One of the most important roles is distinguishing of focused or in many languages also accented syllables from non-accented. For detection of lexical-stress in speech, we are likely to have rather shorter frames of speech for energy extraction, because longer spans could average the sought peaks. When we are looking for stressed syllables, it is useful to measure energy in the middle of the vowels, but this assumes some kind of phoneme time-alignment information.

Similar to micro-intonation, there exists a phenomenon of micro-intensity. The intensity maximum is reached relatively earlier for long vowels compared to the short ones. The inherent/intrinsic intensity causes, that minimum intensity over the speech signal can be observed for final phase of unvoiced occlusive, while the maximum is reached for opened vowels. It is also claimed that prosodic perception of intensity is based on relative proportions [63].

3.1.3 Durational parameters

Subjective length is a perceived result of physical duration. There are various objects for durational examination: pauses, words, syllables, vowels and even all other individual phones. If perceived speech segment length distinguishes between the word meaning, it has phonological function in a language (like vowels in Czech do). In many languages the perceived length plays a role in accent determination. From durational parameters can be derived absolute rates (e.g. speech rate with average syllables or words per second) or they can serve for computation of proportional relative measures. Several intrinsic phenomena can observed regarding the contextual duration of segments [63]. Vowels are shorter before unvoiced obstruents than before voiced (especially in English, probably due to trade-off principle). The duration of vowel also depends on the openness; average duration for Czech [i] is 65 ms while for [a] it is 75 ms (it takes longer to get from a consonant into the open vowel). Here, one needs to distinguish between the segmental and suprasegmental levels: if mentioned [i] and [a] are of the same acoustic duration, than

the [i] will be perceptually considered as longer than [a] [63]. It means that our perception system is used to the intrinsic factors and their neutralizing counterparts are contained within our perception.

3.1.4 Spectral slope

Spectral slope is defined as a ratio of energies in two spectral sub-bands. It is know to be an acoustic correlate of present effort in the vocalization tract [64]. In the figure 3.1 spectrum of two realizations of Czech vowel [a] is illustrated, once with neutral speakers effort (continuous curve), while second time with increased effort. The effort changes the tension in vocal cords influencing the glottal cycle in a way of faster closing which leads to emphasized higher spectral content. Regardless of the fact, that second vocal was globally louder (overall higher curve), it also exhibits an increased magnitude around frequency of 3300 Hz (corresponding to the fourth formant F4).

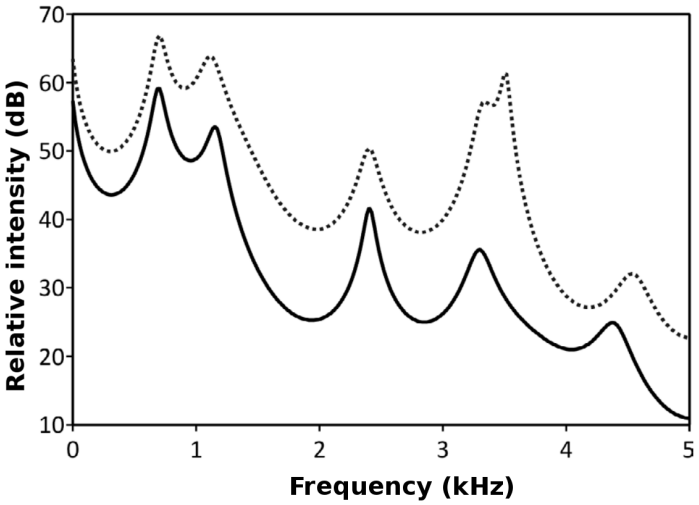


Figure 3.1: Spectral slope difference in two realizations of Czech vowel [a]. A neutral speaker's effort is depicted in full line, while increased effort in dashed line (source [65]).

Spectral-slope in Czech has been lately thoroughly investigated in [66] regarding its sensitivity to speaker, vowel identity and prominence. 10 spectrum related acoustic parameters were evaluated on very "controlled" Czech speech corpus (45 3-syllable pseudowords x 12 male speakers). One-way Anova tests showed significant sensitivity of all tested methods to speaker and vowel identity. Band energy differences with floating pivot of second formant (F2) value was considered as the best method overall for Czech vowels prominence distinction. It was able to differentiate between prominent and non-prominent realization of all Czech short vowels except [u], while keeping the sensitivity to speaker and to vowel identity on reasonable level. The highest speaker discriminative potential in Czech has [e].

3.2 Suprasegmental linear units

The smallest unit of speech that can be marked as prosodic is a syllable. It is the basic domain of suprasegmental phenomenona. Its structure can be seen in the figure 3.2. The mandatory part of each syllable in Czech is the nucleus created by the most sonorant phones - the vowels or syllabic 'r' or 'l' in Czech. There is only one vowel allowed to be present in Czech nucleus, except for diphthongs 'au','ou' and 'eu' (the last one coming from adopted foreign words). The initial consonant (C) part of syllable is called onset (or 'preatura' in Czech), while the concluding consonant part is named coda, both being optional. Multiple consonants are allowed to be present within single onset or coda in Czech, as the example diagram with Czech word 'kost' ('bone' in English) indicates.

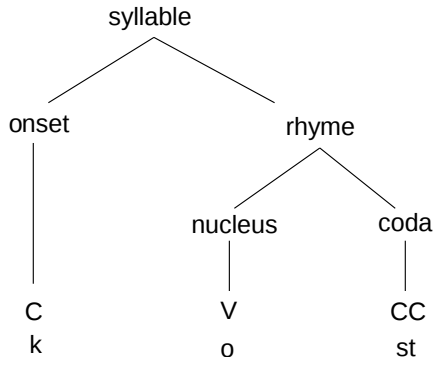


Figure 3.2: Structure of syllable, example for Czech monosyllabic word "kost" ("bone" in English). C stands for consonant, V for vocal. Common division of syllable is into onset, nucleus and coda parts.

Schematic hierarchy of Czech suprasegmental units with real Czech sentence example can be seen in the figure 3.3. In the figure are depicted various levels with multiple terms used to describe the particular level. On the lowest suprasegmental level lie individual syllables that on the lexical level compose words. Unprocessed lexical level, although very suitable for written text, does not have its direct image in the speech in terms of prosody. Its closest suprasegmental level correlates are stress groups (SG, feet) that drive the impression of speech rhythm. Even though several rules exist defining how to segment word sequence into stress units, they do not cover all the possibilities of word combination and generally, this segmentation is often ambiguous and depends on personal preferences of a speaker. One or more stress groups create intonation units, which correspond to the higher complex of information conveyance. Each intonation unit bears the modality information carried by nuclear pitch accent. In Czech, the modality is typically decided on the basis of the very last foot in the phrase bearing the nuclear pitch accent. A sentence is the highest prosodic level and can be composed by more intonation units that create complex or compound sentence. Term utterance very often matches the individual

sentence, but in the ASR field it might also contain multiple sentences spoken by one speaker (for example, between dialogue turn-takes). In this work, utterance will mark the single sentence.

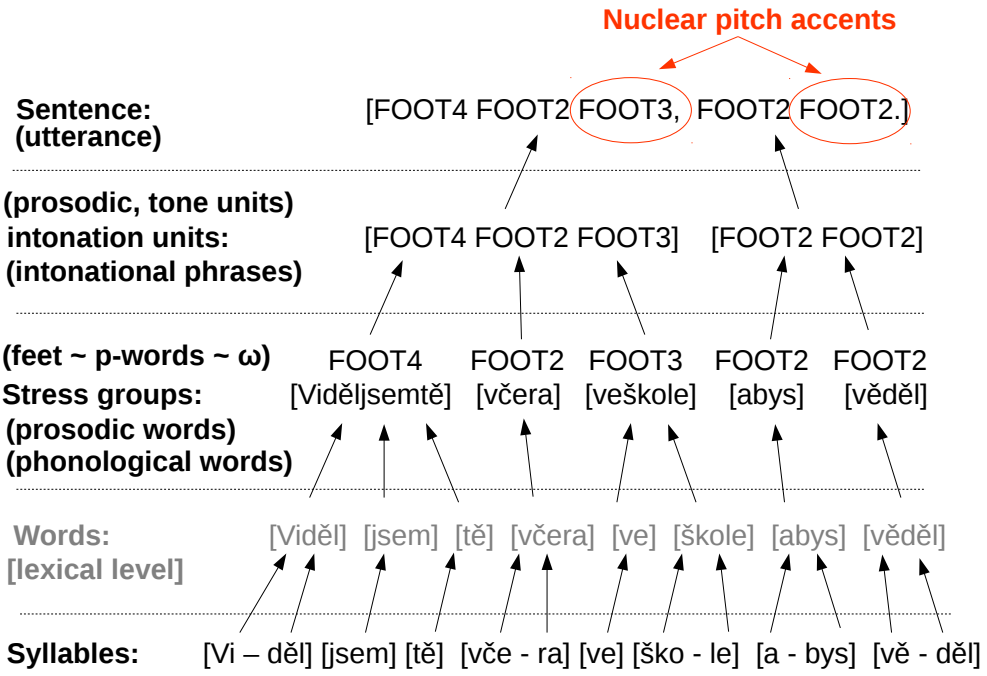


Figure 3.3: Hierarchy of suprasegmental units in Czech (translation of the given sentence is "I saw you yesterday at school, for your information."). The field of Czech clitics is applied in transition from word-level into stress-group level.

3.3 Categorization of Czech as a language

In this section Czech is summarized by various factors (some of them will be covered in the section 3.4 in more detail) and contrast or similarity to other languages, mostly to English, is presented.

Czech belongs to category of Western Slavonic language. While the basic acoustic units of speech are called phonemes, Czech is consistent phonetic language which means that there are obvious rules how certain letters directly correspond to those phonemes. There are about 40 phonemes in Czech (for comparison, English consists of 46 phonemes).

Czech is an inflective language (like the Greek or Latin, compared to English being both agglutinative and inflective) with rich morphology and many inflected word-forms.

Czech is considered to be a syllable-timed language (syllable is considered as basic perceived time resolution step of the speech), while English belongs to stress-timed languages. More information on Czech rhythm can be found in the section 3.4.1.

Czech belongs to the category of fixed stress languages (together with e.g. Polish or Hungarian) with lexical "stress" on the first syllable in the stress-group (this also satisfies Hungarian). Contrarily, English and also Russian (which is together with Czech in Slavonic language family) have hybrid of "free" stress system. English is on its word level similar to Romanian languages (stress is referenced from right edge of the word), while Germanic suffixes (-ing, -ly) do not affect stress and compound words have initial element stress (like in Germanic languages).

3.4 Czech prosody

In this section Czech prosodic system is described in more detail. It initially covers the temporal dimension as rhythmical structure, but main focus is then dedicated to the Czech system of lexical stress and phrase/sentence modality.

3.4.1 Czech speech rhythm

In poetry the rhythm is a basic organizing principle, while in natural speech we find only rhythmical tendencies. Nevertheless, rhythmical structure of languages in natural speech has been separate and growing field of research and Czech is no exception.

Suggestion about rhythmical "schizophrenia" of Czech was presented [67]. Czech exhibits ambiguous behavior in terms of rhythmical categorizations. It fulfills both syllable-timing and stress-timing theorems that are usually sufficient to categorize the rhythm of language. This ambiguity is being explained by sensitivity of Czech to used descriptors for common (and successful) categorization of other languages. This is why multi-dimensional view on languages rhythmical structure is suggested to satisfyingly describe languages as Czech. Czech speech rhythm was revisited in [68], where it supported the previous study and called into question a validity of recent measures for rhythm classification: so called "Grabe & Low model" [69] and "Ramus model" [70].

Regardless the presented research, Czech is still commonly described as syllable-timed.

3.4.2 Czech stress system

Before the sections moves right to the stress system of Czech, few general thoughts are presented at first. The overview of published research follows, with continuation into Czech clitics introduction and concluded by an idea of tone-based model for Czech.

Generalities and terminology

As many studies use inconsistent terminology, the one considered to be the most consistent and relevant to me is the following: Syllabic prominence is achieved by acoustic difference (contrast) in one or in the complex of acoustic descriptors compared to the context (surrounding syllables). Lexical stress is a canonical property of syllable in stress languages and denotes the binary ability of syllable to carry the stress in certain syllabic positions of word (a priori). Acoustically realized stress is called an accent and is carried out (or not carried out) unconsciously. In the spontaneous, but even in read speech, not all the stressed syllables have to be accented. This phenomenon was marked by Peter Auer's claim that the task of prosody is not to portray the syntax. To get the claim into the correct level of understanding, it is exactly the lexical stress that is often denoted as syntactic layer. To fulfill the last piece of terminology mess, the conscious prominence of particular individual word (or less often syllable) is denoted as a focus. The focus does not have to be (and often is not) realized by the same acoustic resources as the accent.

Accent can besides its stress-group level realized also on the highest sentence intonation level, where it is called a nuclear accent and decides about phrase or sentence modality (being mostly part of the very last foot of the sentence). If the accent is not realized on the highest level, it is called pre-nuclear accent.

Known behavior of Czech stress

The lexical stress in Czech is considered as weak compared to other fixed-stress languages No apparent reduction (neither in acoustic quality, nor in its length) of vowels even in unstressed syllables. The absence of length reduction is present due to the phonological function of vowel length as it influences the meaning of the word, e.g. 'vola' as and 'volá').

Stress-group (foot) is the speech unit connected with realized (or non-realized) lexical stress. In the prosodic hierarchy, the stress-group level is higher than syllable level, but lower than phrase level. Stress-group is defined as a group of syllables with at least one stressed syllable. In Czech, only one accented syllable is allowed within each stress-group. and it allows to segment the speech into units that very often correspond to individual words. The causes of those differences are presence of clitics in language (monosyllabic words that lose their independence and create new stress-group with neighboring word) and phenomenon of multi-word stress groups (two lexical words join into one stress-group typically thanks to the high speech rate).

Here, an overview of qualitative and quantitative (in terms of statistical significance testing) research related to Czech lexical stress is presented, while the quantitative research in terms of bigger data volumes and machine-learning oriented experiments on

stress-groups or lexical stress in Czech is in more detail covered in section 6.1.

Although most of native Czech speakers perceive the difference of stressed/unstressed syllables both in production and perception, they are not able to describe the acoustic realizations of this phenomenon. And if so, they often tend to mark stressed syllables higher both in pitch and intensity (which is not true). That is why various experiments have been conducted to find out the objective information on Czech lexical stress.

The very first experiments on Czech stress perception was done in [71]. Perceptual listening tests using fully controlled, delexicalized synthetic speech were carried. All three prosodic parameters (intonation, duration, intensity) are likely to trigger the impression of accent in Czech, but in different proportions. He found out, that typical pitch increase of Czech "stressed" syllable is 1 ST and typical decrease 4 ST. The results are not generalisable because the material does not correspond with the accents present in natural speech.

Later, it is claimed, that the syllable to the left of the accented one, being the last of the preceding stress-unit, is responsible for most durational, dynamic and intonational contrast in the Czech language [72]. This confirms the contextual hypothesis of Czech stress, which emphasizes global prosodic configurations of stress domains, as well as prosodic discontinuities between them. In [72], it is also argued that Czech intonation is resistant to modeling by means of local events, and that the relevant building stone is the stress unit with its holistic contour [73].

In book [74], which is often considered as a "bible" of Czech phonetics and phonology, it is claimed that current explication of lexical stress/accent by acoustic qualities is not satisfactory enough. Impression of prominence is dependent on whole complex of acoustic qualities (pitch, intensity, duration) with the importance on syllable contrast rather than on some absolute values. Nevertheless, the basic difficulty is to describe this complex of acoustic qualities and contrasts specifically. The amount of options that create the impression of the accent is still failing to be generalized enough. Thus, there is still not present any description, which would be able (based on pure acoustic analysis) to decompose continuous text[[speech]] into stress-groups as the Czech listener does. Lexical stress can be realized by either increase or decrease of pitch.

Longer syllable chains within stress group denote even smaller pitch changes in the order of quarter-tones [72]. In spontaneous speech the pitch changes are probably most common features implementing accent.

Later, an investigation of intonation inside stress-groups (feet) using ambiguous syl-

lable chains (real words) in terms of stress-group division was carried out [75]. In Czech, no regular pattern in prosodic properties bound to the stressed syllable has been proven. The more realistic approach seems to be to base the unit delimitation on the linear course of the overall sound properties throughout the unit. The main founding is that pitch contour with clear 'V' shape (decrease-increase contour, drop in the middle) does not implement valid stress-group in Czech and supports the perception of feet delimitation.

In the publication [20] is described the first model for pre-nuclear intonation in terms of various intonation contours depending on the position of stress-group within a phrase or sentence. Its primary aim was an implementation of naturally sounding Czech TTS system.

Helena Spilková in her diploma thesis [76] examined intonation of monosyllabic words on artificially composed ambiguous 1/multi stress-group texts (they can be found in the appendix C as they are used later as testing set in presented experiments). She found out that the stressed syllable in multi-syllable stress-groups is mostly lower than surrounding non-stressed syllables.

Volín in [77] used a statistical approach to investigate the F0 difference between the first and the second syllable within stress-groups. He used read prepared news text by professional speakers (Czech Broadcast 2 news), 1 man, 1 woman, 144 breath groups in total. Careful pronunciation and articulation was expected in this kind of speech, because no visual information for the listener is available during the radio broadcast. Examination of pitch difference between the first and the second syllable in initial, post-initial and final stress-group of each breath group was carried out. The pitch rise between those two syllables has been proven to be statistically significant in all of examined stress-group positions (even in the final one). Nevertheless, out of total 402 stress-groups there were still 20% of them with falling pitch between first and second syllable In terms of size of pitch steps, in the initial stress-group there was average rise of 4 semitones between first and second syllable, while the second and last stress group had 2, respectively 2.3 semitones rise. Anova and Tukey's pair test confirmed significant statistical difference between initial and other two stress-group positions. Two-way Anova test of pitch change with stress-group position and syllable count within stress-group taken as factors: both factors were found to be significant, while also their interaction was significant, which denotes that average pitch differences between stressed and non-stressed syllable do differ according to the stress-group position within the breath-group and also according to the stress-group length. 2-syllable stress-groups have remarkably lower pitch difference than

three and more-syllable stress-groups. Even in their final position, there is clear 2 ST pitch drop in average (because of the finalizing melodeme).

Work [66] was already described in section 3.1.4, nevertheless due to its importance, the main results are summarized again in this place. The ability of lexical stress (prominence) discrimination potential of spectral slope (tilt, balance) was studied. Band energy difference with floating pivot of second formant (F2) value was considered as the best method overall for Czech vowels prominence distinction. It was able to differentiate between prominent and non-prominent realization of all Czech short vowels except /u/, while keeping the sensitivity to speaker and to vowel identity on reasonable level. On the basis of quoted study, the spectral slope was considered as one of acoustic features suitable for the task of stress-group detection in Czech.

The current approach in describing Czech pre-nuclear accent is that the native Czech do perceive the incrementally expanding contours of acoustic descriptors after canonical "stressed" syllable, rather than one-time stress indication [65].

In correlation to the investigated acoustic qualities together with the sense about Czech stress for non-native speakers (who claim to perceive the accent on different than the first syllable of stress-group), there was a suggestion to change in Czech phonetics terminology. The term "přízvukový takt" (a stress-group) was suggested to become "mluvní takt" (which can be translated as a speech-group) in [78].

Clitics in Czech

Clitics are monosyllabic words that have strong tendencies to join preceding stress-group (then denoted as *enclitics*) or join the following stress-group and take the role of canonical stressed syllable of newly raised stress-group (then denoted as *proclitics*). They are mainly responsible for the non-exact matching between lexical level of words and prosodic level of stress-group (see figure 3.3 with prosodic hierarchy of Czech).

A domain of clitics is in more detail covered in the section 6.2, mainly according to the rules stated in [74]. Nevertheless, the rules do usually not cover all the situations that might arise in real utterances and even the rules do not have to be strictly followed by the speaker in reality. To the best of my knowledge, the most current research regarding the monosyllabic prepositions in Czech was done in [79]. It presents an investigation of factors that influence the accent placement on the preposition or to following word (where "violations" from the "proclitical" behavior of Czech prepositions occurs). Among significant factors belong the length of the following word (longer words tend to bear the accent) and a part of speech of the following word (adjectives tend to bear the accent

significantly more often than nouns). The most important factor was confirmed to be the tendency for creation of more rhythmically balanced and regularly ordered stress-groups.

Tone-based models for Czech

Existence of flat pitch accent in Czech was confirmed in [80]. It is a unique situation, when there was firstly verified a presence of such kind of the accent by listening test, but followingly there were no known cues found to correspond with the accent. Tested cues comprised of acoustic, syntactic and positional connections.

From a theoretical viewpoint, the current lack of tone-based models [[for Czech]] might be more than an effect of historical inertia (most traditional approaches of intonation have been contour-based, while discrete targets are a relatively recent innovation).

An attempt to create the first inventory of Czech pre-nuclear pitch accents for read speech supported by auto-segmental theory was done in [81], where the discovered flat pitch accent was included. The inventory can be understood as a set of labels for prosodic stylization of Czech read speech and follows the initial research, where tonally stylized pitch approach can be as successful as common contour based approach [73].

3.4.3 Czech sentence intonation and nuclear pitch accents

This subsection belongs to phrase or sentence modality field in Czech and so deals with speech melody on the highest, phrase/sentence level.

In Czech, there are valid global tendencies across whole sentence, common also in other languages: there is a presence of downtrends in overall sentence melody (measurable F0 declination) and also in durational features (final or pre-boundary lengthening).

The Czech versus English intonation on the sentence/phrase level was thoroughly compared in [82] using material of spoken parallel scripted and non-scripted dialogues in both languages. Next to the observation of key differences between languages, several results interesting for Czech on its own were presented. The key subject of study consisted in examination of length of the tone-unit, position of nucleus in a tone unit, word class functions of nucleus bearers, “FSP” (based on theory of Functional Sentence Perspective) functions of nucleus bearer and pitch patterns of nuclei.

Average length of the tone-unit in non-scripted dialogues was 4.2 words in both languages, while scripted text claimed to lower Czech average to 4.0 (while raising English to 4.7) words.

Position of intonation nucleus on the last word of the tone unit is the most frequent in both languages but being higher in Czech (72-82% in Czech versus 66-69% in English).

The most dynamic elements have a strong tendency to occur in the final position in Czech.

Regarding the word classes of intonation nucleus bearers, tendencies are very similar for both languages: nouns (29-38%) and verbs (22-33%). In non-scripted dialogues is nucleus often beared by interjections and particles (16-18%).

Pitch patterns were divided between falling and rising, while the remaining types presented very low frequencies in both languages (particularly under 12% in Czech). Their statistics were taken for declarative terminal, declarative non-terminal and "Wh-" and "Yes/No" questions. As non-differing in both languages were marked declarative terminal tone units (falling nucleus in 64-84% of cases) and "Wh-"questions (falling nucleus 61-87%). Non-terminal declarative sentences did differ with rising nucleus for Czech in 41-54% of cases and falling nucleus for English with 51-56% of cases. The most important seems to me the difference of "Yes/No" questions: in Czech they are represented by rising nucleus in 41-75% of cases together with falling one in 12-22% of cases, while in English an even ratio was observed (rises 36-51% and falls 36-45%). From this perspective, Czech tends to have more rising intonation both in non-terminal declarative phrases and mainly in Yes/No questions than English.

Palkova in [74] uses two-level approach for sentence melody and modality description. Her terminology is as follows: Czech term for nuclear pitch accent as functional-level intonation unit is called 'melodeme'. Each cadence may have various phonetic realizations and the term for this specific pitch pattern is called 'cadence'. Palkova distinguishes between three nuclear pitch accents:

- concluding descending (M1) that covers declarative sentences, Wh-questions and imperative sentences,
- concluding ascending (M2) used only for Yes/No questions and
- non-concluding or progredient (M3), denoting that sentence as a complex will continue.

Each nuclear pitch accents can have in reality a shape of slightly different pitch contour, basic variants were are also categorized. The conclusion is that even in theory the distinction between nuclear pitch accent can be obvious, in real speech there are sometimes very small differences between them and depend mostly on height of the intonation steps. Particularly this occurs in pair of descending semi-cadence and signed conclusive cadence [74].

Those issues were later studied more deeply in [83]. modality intonation patterns in spontaneous speech from real life dialogues bring certain problems in perception when

isolated from context. Listening test was performed on isolated phrases from acted dialogues. The conclusions is, that in speech with a certain degree of expressiveness and without the support of broader context, a contour of another melodeme (realization of declarative or non-final utterances) may be understood as a question [83].

The complete description of Czech intonation categorization based on tonal approach for both syntactic and semantic levels brings an article [84].

3.5 Prosody annotation systems

An existence of prosodic annotation systems and data labeled according to them can be a tempting shortcut for obtaining input data for further machine processing. This approach of prosody utilization from labeled data is marked as indirect prosody modeling via intermediate abstract phonological categories (as opposed to direct modeling using the prosodic acoustic features of signal). This is why I try to very briefly cover the description of existing prosody annotation systems, while bringing also claimed criticism.

Historically, the ToBI [50] (Tones and Break Indices) was the first proposed system consisting of set of conventions for transcribing and annotating the prosody of speech.

INTSINT was originally developed by Daniel Hirst in his thesis as a prosodic equivalent of the International Phonetic Alphabet (IPA). INTSINT alphabet was subsequently used as labeling scheme in respected book comparing intonation systems of 20 languages [85].

IViE raised as a comparative transcription system for intonational variation in English [86]. While IVTS stands for the transcription system for Intonational Variation, it was derived from IViE but targeted for French [87]. Compared to IViE, it adds an auditory phonetic transcription tier.

RaP stands for Rhythm and Pitch developed by Dilley and Brown in 2005 and described in [88]. As a scheme of labeling the rhythm and relative pitch of spoken English, it was designed to address certain recognized weaknesses of ToBI (e.g., greater phonetic transparency, greater ease of learning and use, capability of labeling multiple levels of prominence, etc.). The results on inter-labeler consistency presented in [89] suggest that RaP is a more than valid alternative to ToBI suitable for a variety of speech prosody research and technology applications.

Although presented variety of labeling schemes, only the ToBI system is described in more detail as its labels are often valid in derived or improved annotation schemes and the knowledge of ToBI labels is very often expected anyway.

3.5.1 ToBI

ToBI prosody annotation standard [50] was created in 1991 as need for the ability of prosodic labeling in standardized way so that researchers and commercial sphere could share the same data.

Not all the researchers do like full ToBI framework, because some studies reported [90] that manual transcription is time-consuming in terms that it can take up 100-200 times the real time depending of the annotator's level. That is why only limited amount of publicly available ToBI labeled databases is available.

Although ToBI as a general prosody annotation framework was primarily designed for American English annotation, it has been fully adopted for various languages so far. Among framework fully integrated language belong English (covers Mainstream American, Southern British RP and Australian varieties), Japanese (standard Tokyo variety), Korean (standard Seoul variety) and GToBI standing for German standard variety. There is also study presenting ToBI framework for Slovak language, the new scheme is called SK-ToBI [91] introducing one new intonational label and canceling other types reputedly not contained in Slovak speech.

ToBI standard description and its shortcomings

There are two main tasks when annotating with ToBI standard: intonation and phrasing transcription.

For intonation transcription, two goals have to be met - "meaning" of intonational events (prominent, continuation, final, etc.) and the shape of the intonational contour. To meet both of them, Pierrehumbert's notation was adopted. It describes intonation as series of pitch accents and boundary tones, each of them can be high (H) or low (L). The pitch accents are then described by appending star sign (*) whereas tones are marked by percentage sign (%, a boundary tone) or minus sign (-, a phrase accent). After, this notation was broadened with the exclamation mark sign (!) denoting downstepping and by special "HiF0" label used for marking exact F0 peak location.

Regarding the phrasing annotation, these are marked by "break indices" numbered scale (formerly 7 levels, later reduced to 5 levels). "0" denotes a cliticized boundary, "1" is used for default prosodic word boundary, "2" marks a boundary between perceived word groups within an intermediate phrase, "3" is used for an intermediate boundary (the one terminated by phrase accent) and finally "4" marks an intonational phrase boundary (terminated by both phrase accent and boundary tone).

The study [92] explains that ToBI standard contains several flaws that have limited

its acceptance and application it was designed for. The evolution of break indices resulted in move away from the perceptual experience of listener (and this is exactly what was the initial aim - to label what we hear). Also, a description of phrasing labels began to include reference to the intonational events such as boundary tones. This created a linkage between the phrasal and intonational tiers of the transcription, which de-emphasized perceptual experience of the listener. Several studies were also made to evaluate how various transcribers can agree on the same annotated material. Even for specialists of the same speech laboratory, their agreement on the specific edge tone label failed to exceed 50% for six of the nine label types. These results together with slow annotation process are unfortunately in direct conflict with the needs of commercial systems.

3.5.2 Conclusion

As a result of this section, most of presented labels in annotation systems cannot thoroughly capture the fine nuances of pitch contours, that are still perceivable. This is why some of the phoneticians refuse to operate on such annotated data if they should serve as a source of deeper analysis. From my point of view, even if there was a ToBI or related annotation system for Czech and if there also existed a Czech speech database annotated with it, I would be very cautious in using its labels as the only input to any machine learning oriented task. Rather, direct processing of acoustic data seem to be the most relevant way, where no information cannot be lost or misinterpreted.

3.6 Stylization of prosodic contours

Stylization of prosodic contours seems to be a better alternative to previously presented prosody annotating systems for deeper analysis because it relates closely to the psychoacoustics and just noticeable differences (JNDs).

The comprehensive and respected study on pitch contour stylization can be found in [7]. While it covers mostly English and Dutch, it summarizes two basic principles in pitch stylization: the continuous approach and syllabic tone approach.

The continuous curve approach is based on our perception of pitch contour as one continuous curve, which is delimited only by phrase boundaries represented by clear pauses. This is also where the “contour stylization” term comes from as the basic principle here is that it is tried to simplify the original contour up to the point when it is as simple as possible, while the stylized acoustic properties are still perceived the same as in the original contour. This approach leads typically into the description of continuous straight lines with limited number of turning points, where perceptually important facts do happen.

The second approach is based on the idea that only syllabic rhyme represents the perceived part of pitch information and the whole contour can be described by discrete syllabic tones. For tones shorter than 100 ms, it is suggested to represent the tone by single level tone value. On the other hand, for tones longer than 100 ms and if they cross the glissando threshold at the same time (see section for more details), such tone should be represented by a contour tone. Considering the particular syllable part which is responsible for the perceived pitch information convey, Hermes shares the founding, that the syllable onset and the first 20-30 ms of the vowel do very little contribute to the overall pitch perception (in those places of the syllable there are lot of transitions in the signal leading to spectrally unstable content which was found to not be perceived as valid pitch information). It is claimed, that it is the moment between 20-30 ms after syllable onset, where the perceived pitch information is encoded. For typical short vowel with length of 60 ms it is about its center. This information is supported by citing the work [93], where although the continuous stylized pitch contour was extracted, the author explicitly mentions, that only pitch information spanning the syllable nucleus is taken into account in his classification algorithms. The importance of perception in this syllable part is supported by the typical intensity peak that can be found here, or only slight intensity decrease after that peak, which can occur little earlier, but remains quite high.

Hermes also notes the unimportance of segmental (phone) level details (micro-intonation, micro-intensity), which do not influence the final prosodic contour and can be eliminated when studying perceptual phenomena.

To conclude work [7], Hermes suggest that it makes very little sense to distinguish between pitch targets and pitch movements as they are both different representations of the same perceived fact (pitch jumps between individual syllabic tones are as important as possible pitch movements for long vowels exceeding the glissando threshold). Next, if pitch contour is represented by continuous lines with turning points, the complete information of perceived pitch is not contained until rhythmical structure of utterance (typically defined by syllable onsets or nuclei centers points) is added.

In respect to this thesis, where pitch information is expected to play one of the key roles, while it tries to mimic human perception by machine learning methods, it was decided to use syllabic tone approach with pitch information extracted in syllable nuclei centers with support of [7] and [93].

3.7 Relationship of speech and music

There are significant parallels between music and spoken languages [94]. Already in 1775, the melody of English at that time was examined and compared to music [95] and it was the very first time when proposal of prosodic notation was presented. Despite the current view on the work is rather conservative and its content lacks the distinguishing between phonology and phonetics [96], it revealed several facts that are applicable even in our age.

Work [97] deals with research oriented on consonant (in terms of pleasant) judgments of musical intervals in musical scales. It investigates the phenomenon using statistical analysis of sound spectrum. In work [98] it was found that formant structure of speech do influence preference of our 12 ST musical scale used as universum in any western music. Another interesting research regarding the relations between music and speech can be found in [99] and [100].

3.7.1 A hypothesis to be tested

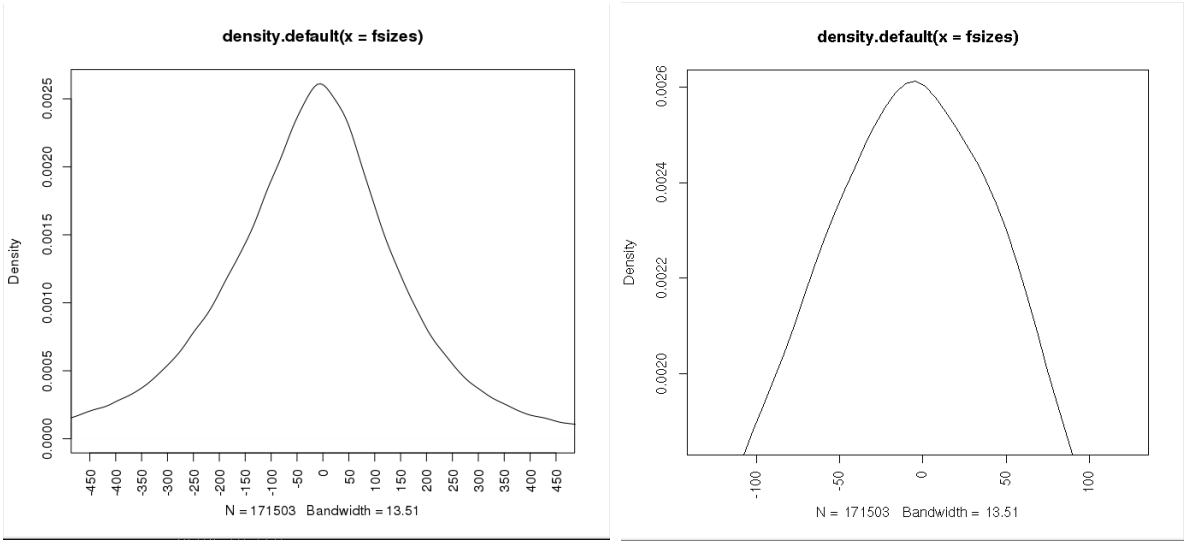
There is undoubtedly a connection between speech melody and musical perception/feeling of a given culture. The semitone based western music needs to some extent project into the melody of speech of those western cultures. During the manual visual and listening verification and investigations of Czech utterances (with pitch extracted at syllable nuclei centers), many times the situation, when almost exact semitone relative steps (and sometimes more in a row) could be heard and visually observed, was faced. The hypothesis to be tested here is to which extent this empirically observed musical semitone relationship between neighboring syllables corresponds to general Czech intonation in speech.

The expectation was that the surrounding around one semitone pitch difference will be more preferred by the speakers for their pitch rising tone-level steps, which would lead into a peak in the histogram (or density function) of all the relative pitch differences in used dataset.

Two datasets were used for examination (see chapter 5 for details): SPEECON "full" subset with force-aligned and algorithmically shifted nuclei centers and its subset with manually labeled nuclei centers (almost 200 utterances).

Resulting densities of pitch difference in cents for the full SPEECON subset is in the figure 3.4. The majority of the function can be seen in the its left part 3.4a capturing the difference range up to ± 4.5 ST, while in its right part a zoom int the range ± 1 ST (100 cents) is depicted.

The suggested hypothesis cannot be confirmed as there is no clear positive peak in histogram around value +100 (corresponds to 1 semitone). Rather, it seems that speak-



(a) Full density (b) Zoomed density

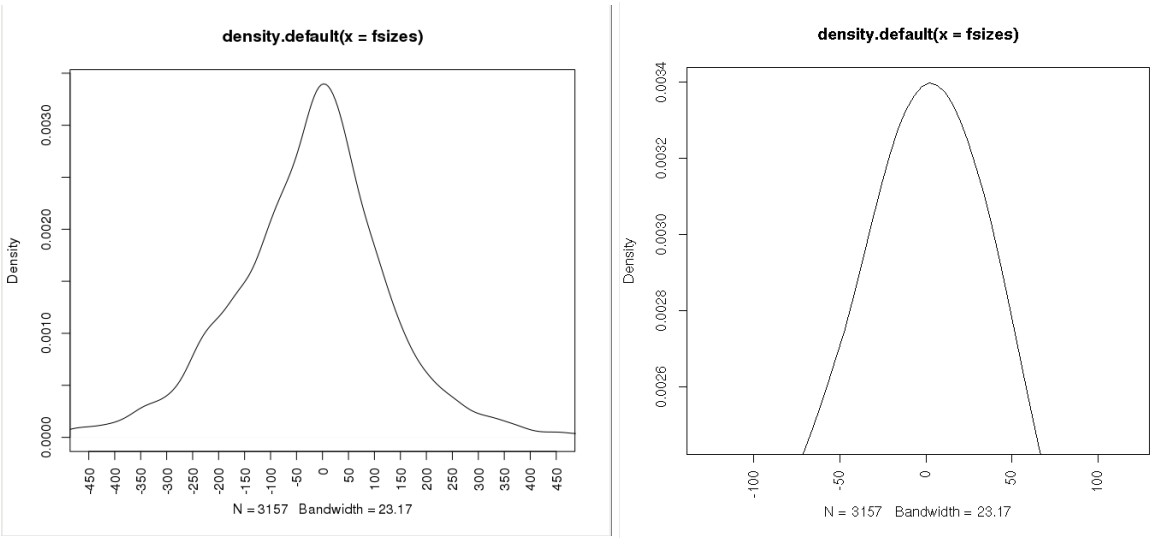
Figure 3.4: Density of relative pitch differences in musical cents on 10062 Czech SPEECON utterances read by 491 speakers (171k of pitch differences in total)

ers have in average almost Gaussian distribution of neighboring pitch differences in their speech. Anyway, slight "bump" in the density function shape can be observed around the value of 40 cents, which suggests very slight preference of pitch increases near the musical interval of half-semitone (a quarter-tone, micro-tonal music basic unit). On the other hand, it can also be seen that the global peak of density function is very slightly negative, which suggests, that most common relative pitch difference is not to stay on the same pitch, but rather make a very slight fall (but this kind of fall will almost surely not be perceived compared to pitch JND).

The measurement was repeated in manually verified (but much smaller) subset of utterances with almost identical results (see figure 3.5). Due to the smaller count of data used for statistics, the overall shape is further from Gaussian, but there is suggestion about bigger and more often falls compared to pitch rises due to the asymmetricity of overall density curve (figure 3.5a). The zoomed version 3.5b does not carry any interesting information this time.

The conclusion here is, that particular speakers might still prefer the pitch difference intervals close to musical universals, but nor the averaged data of 10 062 utterances from 491 speakers neither their manually verified subset did not support the hypothesis.

The experiment might be extended with usage of the only first stress-group for examination (because those are the stress-groups where the behavior was observed most often) and also based on individual speaker data. It is possible that in the given configuration



(a) Full density

(b) Zoomed density

Figure 3.5: Density of relative pitch differences in musical cents on manually verified 189 Czech SPEECON utterances read by 153 speakers (3k of pitch differences in total)

the peak in the distribution around +1 semitone value can be found.

Chapter 4

Modifications of Pitch Detection Algorithms

This chapter is dedicated to field of fundamental frequency (F0) estimation from digital audio signal. Commonly used technical term "PDA" stands for pitch detection algorithm, frequently used synonyms are F0 tracker, pitch extraction algorithm or pitch estimation algorithm. I will stay conform with the most common term PDA, nevertheless it is known that in phonetics or psycho-acoustics the term pitch is rather dedicated to the perceived melody or intonation of speech or generally "height" any sound signal. I believe that speech melody plays a key part among all prosodic acoustic qualities and thus reliable F0 estimation is crucial for further prosodic analysis.

Firstly, general overview is presented followed by description of known PDA methods. The work continues with details on PDA evaluation or comparison with suggestions on extended evaluation criteria and evaluation of known methods itself. The chapter is concluded with several PDA modifications, that seem to be beneficial for pitch detection task whether for speech signals or for any voiced sound signal in general.

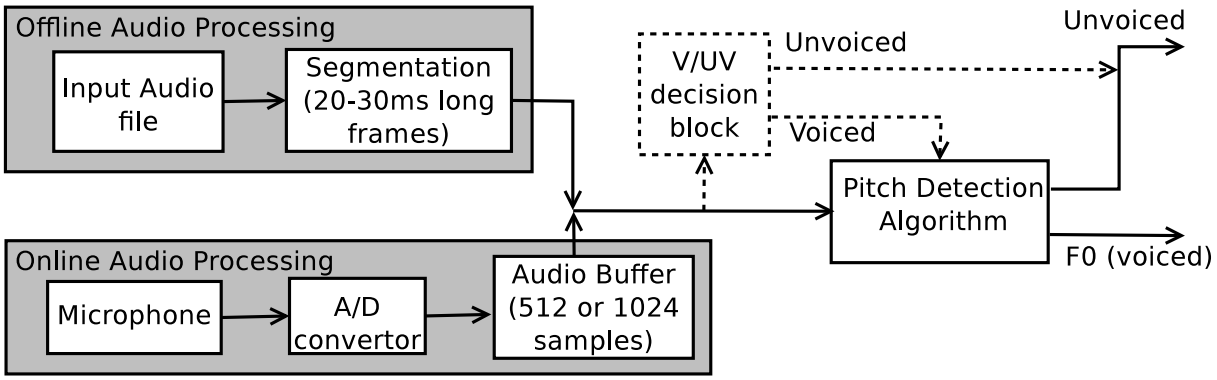


Figure 4.1: Sound processing block diagram for pitch detection algorithm

In the figure 4.1 is depicted typical flow of sound signal that is to be analyzed by the pitch detection algorithm. According to the scenario of usage, two branches are illustrated. The first one is for offline audio processing, where whole audio file is available at the moment of pitch estimation and typically does not invoke any needs on system latency. Thus, longer audio segments with the length about 25 ms can be processed. On the other hand, for real-time scenario using microphone and sound card buffers, whole audio signal is not available at the start of the estimation and the audio samples come incrementally in the chunks specified by the sound card buffer size. This use-case, called often "online" processing, might invoke special demands on low-latency computations implicitly limiting the amount of data available for reliable estimates. Nevertheless, regardless of the signal path, one frame of digitally sampled signal finally reaches the unification point before entering the pitch estimation stage. Then it is being processed and on the output the information about frame voicing and F0 estimate in Hertz unit for voiced frames is expected. If certain PDA is not capable of voiced/unvoiced decision by itself, the optional V/UV block can be pre-ordered in the chain, while only frames classified as voiced are passed into pitch detection block.

4.1 Overview of existing PDAs

There are few well-known PDAs widely used [5]. Their list with short description follows.

In time domain, very simple method based on zero-crossing rate (ZCR) of signal can be used, but unfortunately it is not suitable for complex signals such as continuous speech and therefore is not mentioned further.

The rest of the methods operating in the time domain use the similarity of signal periods by testing various time lags for finding the best correlation and thus need at least two whole periods per speech frame for detecting the lowest desired frequency. These methods are rather robust against additive noise.

Autocorrelation computed in time domain (ACF_{time}, Eq. (4.1)) was the most common PDA in the past, but it is not not sufficient for rapid changes in F0. The pitch period corresponds to maximum of the function. The method is unfortunately prone to octave errors.

$$ACF_{time}(\tau) = \frac{1}{N} \sum_{n=0}^{N-n-1} x(n)x(n+\tau) \tag{4.1}$$

Average Magnitude Difference Function (AMDF, Eq. (4.2)) originated as calculation speed improvement compared to autocorrelation in time domain. The savings can be

done by subtraction instead of multiplication. We are looking for the minimum of AMDF corresponding to the pitch period lag. Method is also prone to octave errors.

$$AMDF(\tau) = \frac{1}{N}\sum_{n=0}^{N-1}|x(n) - x(n+\tau)| \tag{4.2}$$

Combining of methods is quite common (e.g. using complementarity of ACF and AMDF [101]). Shortcomings of autocorrelation method should solve normalized cross-correlation function (NCCF) [102] and is used in many real-time applications. Computationally efficient time domain pitch detection technique Direct Frequency Estimation (DFE) [103] was also presented in the past.

The methods operating in frequency domain (or close-knit to it) utilize the presence of fundamental frequency content replicas (harmonics) in spectrum. That is why strong and regular harmonic content of the signal is needed for these method to work properly.

Autocorrelation in frequency domain(ACF_{freq}, Eq. (4.3)) is direct application of Wiener-Khinchin theorem denoting relation between autocorrelation and power spectrum. Method should give exactly the same results as the time-domain version and suffers from the same shortcomings. The advantage of it is possible calculation speed-up when Fast Fourier Transformation (FFT) is used to obtain frequency spectrum.

$$ACF_{freq}(k) = IDFT\{|DFT(x(n))|^2\} \tag{4.3}$$

Cepstral method (Ceps, Eq. (4.4)) is based on computation of cepstral coefficients, but it might be seen as a non-standard usage of the method compared to standard cepstral coefficient extraction (the main signal spectrum information). The pitch period should be then found as the single peak located right to the main spectrum envelope (Fig. 4.2). Method does not suffer from octave errors, but is showing relatively bad noise robustness.

$$Ceps_k = IDFT\{\ln|abs(FFT(x(n))|\} \tag{4.4}$$

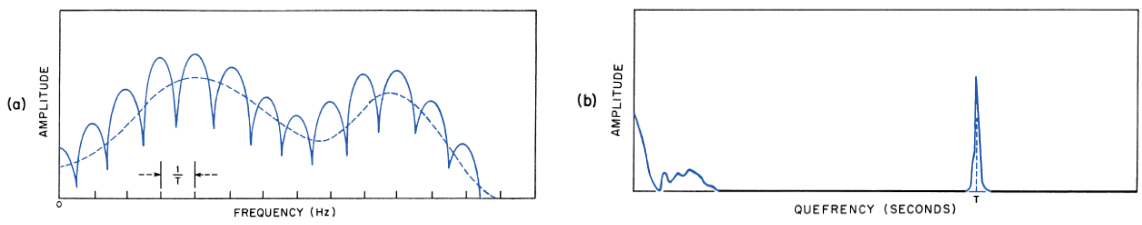


Figure 4.2: Spectrum (a) vs. cepstrum (b), peak at T seconds corresponds to F0 period

4.2 PDAs comparisons with extended criteria

A special framework for pitch detection algorithm evaluation and comparisons was created. Its first version was published in [104]. The result section uses both well-established PDA evaluation criteria and also proposed ones or not widely used so far.

Voicing decision evaluation describes voiced error VE (unvoiced error UE) rate which represents a proportion of voiced (unvoiced) frames that are misclassified as unvoiced (voiced). On the other hand, a precision of the F0 estimate is represented by gross error high GEH (gross error low GEL) criterion. It is a rate of F0 estimates that are correctly classified as voiced but do not meet the 20% upper (lower) tolerance of reference frequency in Hz.

Because of quite large tolerance of GEH and GEL 20% we might not find the differences between two relatively precise PDAs. For this reason, GEH10 and GEL10 criteria are used in analogy to GEH and GEL but with smaller 10% tolerance. They enable us to discover even smaller errors if precision of the estimates really matters. We also use GE as sum of GEH and GEL, and GE10 as sum of GEH10 and GEL10.

Octave errors as special type of gross errors can be often seen in pitch detection. Halving errors (HE) occur when an estimated frequency is near half of reference frequency and doubling errors (DE) are special GEH with estimate being double of reference. We use 1 semitone tolerance around half or double of reference F0 for detecting octave errors.

For global statistical precision comparisons it is suitable to use mean of pitch difference $\overline{\Delta}_{\%}$ (4.5) and standard deviation of pitch differences $\sigma_{\%}$ (4.6), both measured in musical cent units (100 cents equals to 1 semitone) [103].

$$\overline{\Delta}_{\%} = \frac{1200}{N} \sum_{n=1}^{N} \log_2 \frac{F_{est}(n)}{F_{ref}(n)} \tag{4.5}$$

$$\sigma_{\%} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} [1200 \log_2 \frac{F_{est}(n)}{F_{ref}(n)} - \overline{\Delta}_{\%}]^2} \tag{4.6}$$

Regarding the comparisons of precision between two different PDAs, a misinterpretation of the results can occur. Unfortunately not many papers note this phenomena (next to our experiences we found it also in [105]) and reader is probably expected to understand it automatically. It needs to be realized that voicing errors and precision errors are not independent. If voicing error rate (VE) differs too much for given PDAs it is unwise to directly compare their precision by gross errors. With increasing VE criterion it is likely that the PDA will tend to lower its gross error rates because some precision-problematic

frames were probably already marked as non-voiced in VUV stage (and vice-versa). Also, serious PDA precision comparison involves a usage of really the same VUV mechanism.

It is also useful to able to determine gross errors not across the entire frequency band but for example separately within five smaller frequency sub-bands (near 2/3 octave bands were used to cover theoretical speech range up to 500 Hz with some reserve in the highest band). Band boundaries were chosen ad-hoc, but with respect to common male voice range (second band 88-141Hz), common female range (third band 141-225Hz) and higher female range (fourth band 225-353Hz). We follow summed GE and GE10 precision in each of the band with an additional information about correct voicing detection percentage in the particular band.

4.2.1 Used dataset and experimental setup

Before the creation of large ECESS pitch reference database based on part of Spanish SPEECON [106], most of researchers used either Keele Pitch Reference database [107] or CSTR Bagshaw [108] reference databases. There is a study [105] that uses three various reference databases at the same time for evaluation, but provides only cumulate results and on different PDA set that our paper presents. On the other hand, our method is to evaluate each of the databases separately and thus enables the discovery potential relationships between them.

For the evaluation of our experiment we used all of three widely used pitch reference databases. Their detailed descriptions with some useful information follows together with their SNR mean ($\overline{SNR}$) and SNR standard deviation (σ_{SNR}) values computed using the SNR tool [109] that implements segmental SNR estimation and integrated cepstral voice activity detector. Obtained SNR values give us better knowledge about signal strength compared to the background noise in the recordings for each database (or its channels).

First used pitch reference database is manually labeled part of Spanish SPEECON database [106]. The overall length is 60 minutes (1024 utterances recorded by 60 speakers, 30 males and 30 females). Database contains speech signals with reference pitchmarks, derived F0 values (1ms step) and voice activity reference. Pitchmark positions were firstly estimated by pitch marking algorithm [106], all of them were then manually checked by authors. There are 4 channels (ch0-ch3) recorded at different distances from speaker which leads to varying SNR across the channels. Parameters of the audio data are following: sampling frequency (FS) 16kHz, Little-Endian, 16-bit dynamic range. Available categories are: Male/Female, Office, Public place, Car.

Measured SNR values for all 4 channels are:
$\overline{SNR}(ch0) = 26.8$dB, $\sigma_{SNR}(ch0) = 7.7$dB, $\overline{SNR}(ch1) = 11.7$dB, $\sigma_{SNR}(ch1) = 8.7$dB, $\overline{SNR}(ch2) = 11.3$dB, $\sigma_{SNR}(ch2) = 7.4$dB, $\overline{SNR}(ch3) = 5.6$dB, $\sigma_{SNR}(ch3) = 7.8$dB.

We decided to run all the experiments on two channels of SPEECON database with highest SNR (ch0 and ch1), because third channel (ch2) shows, according to our measurements, similar SNR behavior as the second channel (ch1) and SNR values in fourth channel (ch3) are too low for most of wanted applications (special noise reduction preprocessing will probably be needed for serious evaluations).

Secondly, we used Keele Pitch Reference database [107] with speech signals of total length about 5 minutes and recorded by 10 speakers (5 males and 5 females). Each speaker reads phonetically balanced text ("North Wind and the Sun" story, each utterance is about 30s long). Audio data are recorded in one channel and sampled with 20kHz, Little-Endian, 16-bit. Pitch references are obtained from attached and delay-corrected laryngograph signal using autocorrelation function with time step 10ms. Database can be divided by gender category (M/F) only, $\overline{SNR} = 27.5$dB with $\sigma_{SNR} = 5.0$dB.

Thirdly, CSTR Bagshaw pitch reference database [108] was used. It contains 100 utterances (50 sentences read by one male and one female English speakers consisting of 15 questions and 35 statements) in one audio channel (FS=20kHz, Big-Endian, 16-bit). Reference F0 contours were derived from attached laryngograph signal by database authors and are stored in ms-Hz pairs rather then in fixed time step format. Available category is M/F, $\overline{SNR} = 33.5$dB with $\sigma_{SNR} = 6.5$dB.

Except the standard PDA, state-of-the-art pitch estimate algorithm were also tested for completeness. Those algorithms are easily accessible via open-source speech manipulating tools. Wavesurfer tool [110] uses internally Snack Sound Toolkit [111] implementation for pitch detection algorithms, we used its ESPS method with default settings (denoted as $\text{Wavesurfer}_{\text{ESPS}}$ in result section). Praat [112] is another widely used speech tool and we used its auto-correlation (Praat_{ac}), cross-correlation (Praat_{cc}) and subharmonic summation ($\text{Praat}_{\text{shs}}$) methods for the comparisons.

4.3 Modifications of PDA output post-processing

This section brings two main areas of PDA output post-processing modifications. The first one resides in modification of candidates distance measure in transition probability

function for dynamic PDA output post-processing, while the second one deals with an inclusion of time-dimension into this transition probability function.

Transition probability function definition

Optional post-processing stage of PDA tries to smooth the final pitch contour of detected fundamental frequencies and often corresponds to the elimination of octave errors (pitch doubling and halving) or outlier values. This is feasible because speech pitch is assumed not to vary much between adjacent frames. A widely used technique for smoothing is 5-point median filtering with good ability of outliers rejection and also capability of correction of one-frame V/UV misclassification. If there is a possibility to get more F0 candidates from the core PDA function, the dynamic programming methods such as Viterbi algorithm [113] can be used for finding the most probable pitch contour [114].

4.3.1 Transition probability function - distance measure

In this section it is shown the importance of suitable definition of transition probability function in dynamic pitch track post-processing. The section firstly deals with part of Kotnik's (et al.) [115] pitch detection algorithm based on merged normalized backward-forward correlation (MNFBC) and shows the improvements in its precision on female signals (while lowering overall computational requirements) when transition probability function of Viterbi [113] post-processing algorithm is defined correctly for speech signals. For comparison of mentioned method with other well-known and commonly used PDA methods, we used 3 different pitch reference databases respected by researches in the field. Also, next to the ordinary criteria used for PDA evaluation the paper brings precision scores derived from particular 2/3-octave wide frequency bands. All of the obtained results seem to be consistent across all used databases.

Transition probability function between candidates k and l can be defined variously, but an intuitive approach for speech signals is following: due to the logarithmic distribution of musical tones across frequency range we firstly convert candidate distance from frequency scale into musical semitones cents $\Delta F_{k,l}^{(c)}$ in equation (7.1), 1200 cents difference equals to musical octave interval. In [102] similar approach was presented defining conversion for candidate lags instead of frequencies. This is also the step missing in the post-processing algorithm presented in [115]. Finally, transition probability function $a_{k,l}$ can be defined as decreasing exponential of candidate distance in equation (4.8), multiplying constant was inspired by [115] and verified to give best results. Such defined transition

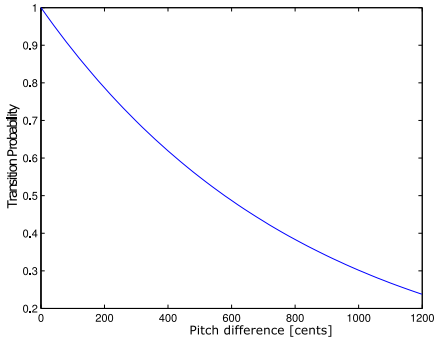


Figure 4.3: Transition probability function of cents difference up to one octave

probability function follows the speech property that speech pitch usually differs very little in neighboring frames. Bigger pitch jumps are allowed, but penalized by lower probability of transition (figure 4.3). Moreover, with Hz to cents conversion there disappears the phenomenon of transition probability decrease in higher frequency bands (corresponding to female signals). Also, an increase in the pitch by certain musical interval is then as probable as the decrease by the same interval.

$$\Delta F_{k,l}^{(c)} = 1200 |\log_2 \frac{F_k}{F_l}| \tag{4.7}$$

$$a_{k,l} = \frac{1}{e^{0.0012\ \Delta F_{k,l}^{(c)}}} \tag{4.8}$$

4.3.2 Tested algorithms

We illustrate the enhancement of transition probability function on the part of the complex PDA described in detail in [115]. It uses merged normalized backward-forward correlation (MNFBC) as its core pitch-detection method. For simplification of the MNFBC formal description we firstly define correlation term in equation (4.9), where x_w denotes a frame of the discrete speech signal and constant MAX_PER refers to time period (in samples) of lowest detectable frequency. Difference between k and l represents the tested lag. Note, that with suggested frame length of $4 * MAX\_PER$ we do not have to worry about crossing the frame boundary in the correlation term. Having this definition, we can compute normalized forward (NFC) and backward (NBC) correlations according to equations (4.10) and (4.11), where t is tested lag. With denominator providing geometric energy mean of the sub-frames (shifted by t samples), the NFC function is generally very similar to normalized cross correlation function (NCCF) defined in [102]. NFC

$$NBC[t] = \frac{< x_{w_{2MAX\_PER}}[n], x_{w_{2MAX\_PER-t}}[n] >}{\sqrt{< x_{w_{2MAX\_PER}}[n], x_{w_{2MAX\_PER}}[n] >< x_{w_{2MAX\_PER-t}}[n], x_{w_{2MAX\_PER-t}}[n] >}} \tag{4.11}$$

$$MNFBC[t] = \frac{< x_{w_0}[n], x_{w_0}[n] > (NFC'[t])^2 + < x_{w_{2MAX\_PER}}[n], x_{w_{2MAX\_PER}}[n] > (NBC'[t])^2}{< x_{w_0}[n], x_{w_0}[n] > + < x_{w_{2MAX\_PER}}[n], x_{w_{2MAX\_PER}}[n] >} \tag{4.12}$$

function operates within the range of first three quarters of the processed prolonged frame, while NBC operates on last three quarters. Final MNFBC function is then obtained by combination of half-way rectified NFC' and NBC' functions in (4.12).

$$< x_{w_k}[n], x_{w_l}[n] > = \sum_{n=0}^{2*MAX\_PER-1} x_w[n+k]x_w[n+l] \tag{4.9}$$

$$NFC[t] = \frac{< x_{w_0}[n], x_{w_t}[n] >}{\sqrt{< x_{w_0}[n], x_{w_0}[n] >< x_{w_t}[n], x_{w_t}[n] >}} \tag{4.10}$$

In contrast to [115] we replaced a trained neural network in the voicing decision stage by thresholding the peak values of the core merged normalized forward-backward correlation (MNFBC) function together with flexible energy thresholding of the frame (we firstly compute energy range of current whole utterance). After normalization of core function values so that first MNFBC value equals to one, the core function threshold was set to 0.27 and energy threshold was set to one third of the obtained utterance energy range.

With improvements to transition probability function in dynamic post-processing suggested in this experiment (see section 4.3.2 for results) there is also no need to ensure the correct pitch with sub-harmonic summation (SHS) [116] method to prevent halving errors. In online applications, the prolonged frame with length of $4 * MAX\_PER$ doubles usual delay needed for gathering all the frame samples, which makes the whole concept of MNFBC function not suitable for low-latency real-time processing.

Results and Discussion

Result section consists of three subsections. Together, they should create comprehensive view on the performance of various post-processing methods and chosen pitch detection algorithms. Particularly, we introduce frequency content of used databases, post-processing influence on male and female signals and overall results for various PDAs on reference

Speecon	**Frequency band [Hz]**				
	<88	88-141	141-225	225–353	>353
Abs. [ ]	74603	714924	906385	222798	2678
Rel. [%]	3.88	37.21	47.17	11.60	0.14
Keele	**Frequency band [Hz]**				
	<88	88-141	141-225	225–353	>353
Abs. [ ]	16000	59320	63890	30260	130
Rel. [%]	9.43	34.98	37.67	17.84	0.08
CSTR	**Frequency band [Hz]**				
	<88	88-141	141-225	225–353	>353
Abs. [ ]	2469	43660	25889	51106	118
Rel. [%]	2.00	35.43	21.01	41.47	0.10

Table 4.1: Absolute (Abs.) and relative (%) occurrences of reference F0 in particular frequency bands across 3 tested databases

databases including their behaviour in particular frequency bands.

Frequency Content of Used Databases

Firstly, we focused on reference F0 distribution according to suggested five disjoint frequency bands to see what is the fundamental frequency content of the utterances in databases. Absolute and relative (%) distributions can be found in Tab. 4.1. We expected smallest occurrences in the lowest (bass male range) and in the highest (children range) frequency bands which has been confirmed. Highest ratio of the lowest reference fundamental frequencies can be seen in KEELE pitch database (over 9%), while CSTR Bagshaw database contains highest portion of references in fourth (higher female) band.

Gender-Based Post-Processing Results

In the next part of the experiment, we deal with post-processing options applied on MNFBC pitch-detection algorithm (described in section 4.3.2). We show the differences in transition probability function definition of pitch contour dynamic post-processing separately on male and female signals. In tables 4.2-4.5 the label $\text{MNFBC}_{\text{no\_post}}$ stands for the variant with no post-processing applied, $\text{MNFBC}_{\text{medfilt5}}$ means application of 5-point median filter, $\text{MNFBC}_{\text{Vit\_HzDiff}}$ denotes Viterbi dynamic post-processing using transition probability function based on pure frequency difference and finally, $\text{MNFBC}_{\text{Vit\_CentDiff}}$ marks Viterbi dynamic post-processing with transition probability function based on musical pitch (cents) distance.

Although the same voice/unvoiced decision procedure (and core pitch estimation func-

tion) is used in all tested post-processing methods, we can see that median filter generally modifies the VE and UE rates. This is why we left VE and UE criteria in the comparison tables to illustrate this behaviour. In general, both VE and UE rates are slightly improved with median filtering, which can be explained by the ability of outliers correction. The only exception can be seen in SPEECON channel 0, where median filtering brings little raise in unvoiced errors. From the results we can see that male signals are more prone to voiced errors (VE).

The most important observation is that $\text{MNFBC}_{\text{Vit\_HzDiff}}$ version suffers from high rate of low gross errors on female signals (worst case is almost 59% GEL rate in Bagshaw database), while most of them (around 90% for all the databases) are caused by halving error occurrences. This is in compliance with theory because lower octave jump has much greater transition probability than higher octave jump computed using Hz units. On the other hand, it lowers the male signals summed GE criterion on lower SNR channel 1 of SPEECON database, but this single success cannot compensate overall fail for women signals and generally, $\text{MNFBC}_{\text{Vit\_HzDiff}}$ often brings addition errors compared to version $\text{MNFBC}_{\text{no\_post}}$ without any post-processing done.

Due to the clear arrangement of the male/female comparison tables 4.2-4.5 we omitted gross errors criteria with 10% tolerance range, because they follow (with some additional constant) gross errors with 20% tolerance quite well over the whole set of databases. The only noticeable observation is increasing difference between $\text{MNFBC}_{\text{medfilt5}}$ and $\text{MNFBC}_{\text{Vit\_CentDiff}}$ precision (to the benefit of $\text{MNFBC}_{\text{Vit\_CentDiff}}$) so that cent-difference based Viterbi post-processing wins across all the databases in both gender categories in GE10 criterion (compared to SPEECON channel 1 male and KEELE male signals where 5-point median filtering shows least amount of summed 20% tolerance gross errors GE).

In total, smoothing of pitch contour with Viterbi algorithm using cent difference in transition probability function ($\text{MNFBC}_{\text{Vit\_CentDiff}}$)) shows the best precision scores in our post-processing experiment on male/female categories (Tab. 4.2-4.5). But also, simple 5-point median filtering ($\text{MNFBC}_{\text{medfilt5}}$) can still be a good choice especially on 20% precision tolerance level and when lower computational demands matter.

Global Results

In the last part of the experiment we present overall results in paired tables A.1-A.4 using all evaluation criteria described in section 4.2. We evaluated three PRAAT methods and one Wavesurfer tool method as described at the end of section 4.3.2. For SPEECON database we also used results for Kotnik2009 PDA method taken over from [115]. Al-

SPEECON Ch0 M/F	VE	UE	GEH	GEL	GE	HE	DE
$MNFBC_{no\_post}$	12.87 / 7.83	9.71 / 11.90	1.27 / 0.82	1.84 / 4.42	3.11 / 5.24	1.05 / 3.61	0.09 / 0.14
$MNFBC_{medfilt5}$	11.25 / 6.60	10.02 / 12.61	0.65 / 0.54	1.91 / 4.15	2.56 / 4.69	0.90 / 3.20	0.04 / 0.05
$MNFBC_{Vit\_HzDiff}$	12.87 / 7.83	9.71 / 11.90	0.74 / 0.37	3.39 / 44.39	4.13 / 44.76	2.43 / 40.48	0.11 / 0.04
$MNFBC_{Vit\_CentDiff}$	12.87 / 7.83	9.71 / 11.90	0.92 / 0.69	1.26 / 2.27	2.17 / 2.97	0.49 / 1.78	0.11 / 0.07

Table 4.2: SPEECON Channel 0 - post-processing results on Male/Female signals [%]

SPEECON Ch1 M/F	VE	UE	GEH	GEL	GE	HE	DE
$MNFBC_{no\_post}$	27.63 / 21.53	8.21 / 10.33	19.46 / 10.52	2.75 / 4.83	22.22 / 15.35	2.06 / 3.40	0.16 / 3.70
$MNFBC_{medfilt5}$	26.31 / 20.39	8.13 / 10.52	18.30 / 9.33	2.68 / 4.35	20.98 / 13.68	1.85 / 3.01	0.12 / 3.57
$MNFBC_{Vit\_HzDiff}$	27.63 / 21.53	8.21 / 10.33	15.41 / 4.82	3.23 / 19.7	18.64 / 24.52	2.14 / 16.67	0.67 / 1.49
$MNFBC_{Vit\_CentDiff}$	27.63 / 21.53	8.21 / 10.33	19.89 / 9.17	1.37 / 2.15	21.26 / 11.32	0.70 / 1.39	0.16 / 3.14

Table 4.3: SPEECON Channel 1 - post-processing results on Male/Female signals [%]

KEELE M/F	VE	UE	GEH	GEL	GE	HE	DE
$MNFBC_{no\_post}$	19.62 / 8.25	15.79 / 23.01	1.88 / 1.83	0.51 / 3.17	2.39 / 5.00	0.20 / 2.94	0.83 / 0.89
$MNFBC_{medfilt5}$	18.65 / 6.90	12.85 / 22.56	1.24 / 1.31	0.53 / 2.37	1.77 / 3.69	0.09 / 2.15	0.59 / 0.69
$MNFBC_{Vit\_HzDiff}$	19.62 / 8.25	15.79 / 23.01	1.86 / 1.67	0.91 / 21.65	2.77 / 23.32	0.60 / 21.09	0.90 / 0.83
$MNFBC_{Vit\_CentDiff}$	19.62 / 8.25	15.79 / 23.01	1.93 / 1.90	0.37 / 0.53	2.30 / 2.44	0.06 / 0.40	0.91 / 1.09

Table 4.4: KEELE DB - post-processing results on Male/Female signals [%]

BAGSHAW M/F	VE	UE	GEH	GEL	GE	HE	DE
$MNFBC_{no\_post}$	13.15 / 8.67	19.96 / 17.42	0.96 / 0.95	0.96 / 3.19	1.93 / 4.14	0.09 / 2.37	0.15 / 0.23
$MNFBC_{medfilt5}$	11.32 / 7.79	18.97 / 17.23	0.85 / 0.87	1.25 / 2.13	2.09 / 3.00	0.08 / 1.28	0.13 / 0.17
$MNFBC_{Vit\_HzDiff}$	13.15 / 8.67	19.96 / 17.42	0.92 / 0.42	1.17 / 58.97	2.10 / 59.39	0.31 / 53.59	0.14 / 0.07
$MNFBC_{Vit\_CentDiff}$	13.15 / 8.67	19.96 / 17.42	0.90 / 1.05	0.88 / 1.56	1.78 / 2.61	0.03 / 0.94	0.17 / 0.25

Table 4.5: CSTR BAGSHAW DB - post-processing results on Male/Female signals [%]

though different evaluation scripts were used for the evaluation, the results should be roughly comparable due to similar error rates obtained for Praat algorithms, which were also evaluated in [115] (with respect to the fact, that our evaluation system seems to be slightly more strict in all mutual criteria). We omitted $\text{MNFBC}_{\text{Vit\_HzDiff}}$ algorithm from overall result tables due to its unacceptable behavior on female signals.

Impact of trained neural network on voicing decision is obvious for Kotnik2009 PDA (lowest summed VE+UE), closely followed by $\text{Wavesurfer}_{\text{ESPS}}$ on SPEECON channel 0. Also gross error rates are at lowest level for Kotnik2009 PDA. Except higher rate of unvoiced errors (UE) the results of $\text{MNFBC}_{\text{Vit\_CentDiff}}$ on high SNR signals (SPEECON Channel 0, Keele and Bagshaw databases) are almost comparable to other pitch detectors contained in public speech tools. PDA with best computed average precision error varies with used database (Praat_{ac} for SPEECON channel 0, $\text{Wavesurfer}_{\text{ESPS}}$ for channel 1 and Praat_{ac} for Bagshaw and Keele databases).

Each of overall performance tables also presents comparison of precision in five distinct frequency bands. Used aggregated criteria format "GE/GE10 @ (100-VE)" shows summed precision gross errors in both tolerance ranges (GE/GE10 part) and the ratio of frames in the particular band correctly classified as voiced (100-VE part). This enables us to consider the importance of given precision and also the ability of given PDA to detect voiced frames in particular frequency range.

By comparing the frequency band overall results with male/female results for MNFBC based PDAs, we can assume that there is obvious conjunction between two lowest frequency bands and male signals, and third and fourth frequency band and female signals, as expected. Male/female voiced errors and precisions can be approximately obtained from corresponding frequency bands as weighted averages of traced criteria, where the weights are given by the ratios of F0 distribution in these corresponding frequency bands (see section 4.3.2). Obtained results don't have to be accurate due to an overlay of male and female voices on used 141 Hz boundary (there can be female voices below and also male voices above this threshold).

Generally, the lowest and the highest frequency bands bring more voicing and gross errors than the bands lying in-between. F0 located in the highest band is quite rare, but F0 in the bass male frequency band (below 88Hz) are present in many utterances (almost 10% of Keele DB). We were surprised by worst performance of Praat_{ac} in the lowest band of SPEECON channel 0. Best voicing decision capabilities in this band seems to have

$\text{Wavesurfer}_{\text{ESPS}}$ algorithm (except SPEECON channel 1 where it is the worst) and conversely, our MNFBC based algorithms have often troubles with voicing decision and also precision in the lowest band. It is difficult to do any further conclusions because situation in three middle frequency bands is quite balanced.

Regarding the robustness of tested algorithms we prepared results for signals with lower SNR values (SPEECON channel 1) in the Tab. A.2. We can see that most of tested algorithms are influenced by an expected decrease of voicing decision success rates and precision. Kotnik2009 algorithm has by far the lowest voicing error rates. This is caused by its special noise reduction pre-processing block, which also positively influences precision of the estimates. Praat_{cc} produces best voicing decisions from the rest of PDAs, but is outperformed by $\text{Wavesurfer}_{\text{ESPS}}$ in precisions, which is this time very closely followed by $\text{Praat}_{\text{shs}}$ (except really high UE). $\text{MNFBC}_{\text{Vit\_CentDiff}}$ is for SPEECON channel 1 comparable to Praat_{ac}.

Conclusions

We have explained and verified by the results an importance of suitable transition probability function definition for dynamic pitch post-processing. Results confirm that a computation of transition probabilities with pitch difference measured in semitones (or cents) significantly improves low gross error rates (especially frequency halving) for female signals over the case of direct difference of frequencies. The improvement has been verified on separate male and female signals across all tested reference databases. The biggest impact was observed by the decrease of halving error rate for female signals of Bagshaw database (from 53% to nearly 1%) whilst GEH values increased very reasonably (from 0.4% to 1%). Although this minor change prove to be significant in the right context, the overall experimental PDA did not outperform the complex Kotnik's PDA presented in [115]. But with relative implementational simplicity and spared computational complexity it could be still good choice for some application with high SNR signals, which was also confirmed by results compared to other common tested PDAs. Separate frequency band-based evaluation criteria also brought another useful insight into behaviour of all tested PDAs and surely can help to examine their strengths and weaknesses for particular use-cases.

4.3.3 Transition probability function - temporal forgetting

This experiment loosely follows the experiment presented in the previous section. Viterbi post-processing is reminded again here for completeness .

Viterbi post-processing

Post-processing using the Viterbi algorithm [113] is applied to find optimal track of the pitch. The power of Viterbi procedure lies in its ability to apply user-defined rules for comparing the candidates. However, for proper use of the algorithm some requirements have to be met. Each candidate needs to have assigned its "emission" probability b_k (the probability that candidate k is F0 for the current frame, without considering any history) and also its "transition" probability a_{kl}, which denotes how probable it is that candidate k in the current frame will be followed by candidate l in the next frame. Having these context-independent values we can gradually compute the values of function $\delta_{m,l}$ which tells the final probability of candidate l being F0 for frame m considering the results from the the previous frames. Function $\psi_{m,l}$ designates the index of the most probable candidate in frame $m-1$. The equations can be then expressed as (4.9) and (4.10).

$$\delta[m,l] = \max_k[\delta[m-1,k]a[k,l]]b[l] \tag{4.9}$$

$$\psi[m,l] = \arg\max_k[\delta[m-1,k]a[k,l]] \tag{4.10}$$

The algorithm starts by assigning for the first frame $\delta_{1,i} = b_i$ and $\psi_{1,i} = 0$.

In the current implementation the three best candidates (three highest peaks) come from MNFBC function. Note that these candidates often (but not always) correspond to the harmonic content of a speech signal [8]. This means that in most cases the candidate with highest MNFBC value is really F0, and the two other highest values are harmonics of F0 (its natural multiples). However, there could be also cases when harmonics are "stronger" than fundamental. In this case, the Viterbi procedure should prevent from halving or doubling errors. The basic scheme of the algorithm is depicted in figure 4.4.

Now let us consider a situation when it is possible to have a jump in pitch of speech in the place of the border of neighboring prosodic units [74] (with unvoiced segments between them, so that the first v oiced segment of the new prosodic unit is the next voiced segment for the last prosodic unit in terms of the Viterbi algorithm). The previous probability function will not allow the change to be immediately applied to the pitch track, and needs some time to "adopt" the new pitch level.

Most utterances take place in the range of the musical fourth (which is 5 semitones = 500 cents in terms of explicit musical distance). This is not the overall pitch range, but it is the common range that we use across prosodic units, sometimes referred to in the literature as the "pitch sigma". It is probable that the biggest jump in pitch of 5 semitones will not occur very often and only on the boundaries of prosodic units. However, we permit this jump to be possible without any penalization after a long enough prosodic

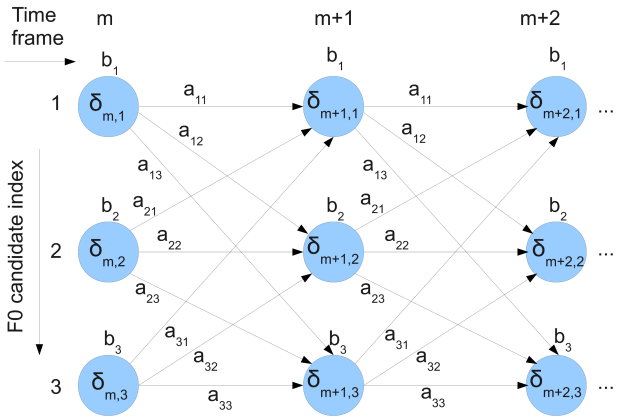


Figure 4.4: The trellis of the Viterbi algorithm used

pause. For this reason, a difference of 500 cents is a limit, and higher differences will be penalized by a linear decrease.

The behaviour of the temporal probability function of two variables (cent difference x and time t) can thus be expressed in the cent difference interval $x \in \langle 0, 500 \rangle$ as:

$$a(x,t) = e^{-0.0012x} + \frac{(1 - e^{-0.0012x})t}{T_{thr}} \tag{4.11}$$

and on the interval $x \in \langle 500, 1200 \rangle$ as:

$$a(x,t) = e^{-0.0012x} + \frac{(\frac{1200-x}{700} - e^{-0.0012x})t}{T_{thr}} \tag{4.12}$$

where T_{thr} is the time forgetting threshold. When the prosodic pause length reaches this value, all pitch changes in the range of 500Hz have a transition probability of 1.

Test conditions

All the results were computed using a manually labeled pitch-reference database as part of Spanish SPEECON [106], with the use of a pitch evaluation framework [117]. All parts of the proposed algorithm were implemented in MATLAB environment.

4.3.4 Results and discussion

Table 4.6 shows the overall results for the highest signal-to-noise (SNR) ratio channel 0 of the reference database. MNBFCv1 is the basic variant with the voiced/unvoiced (V/UV) decision threshold set to value 0.5 and with the transition probability of the Viterbi procedure computed from the direct frequency difference. MNBFCv2 improves

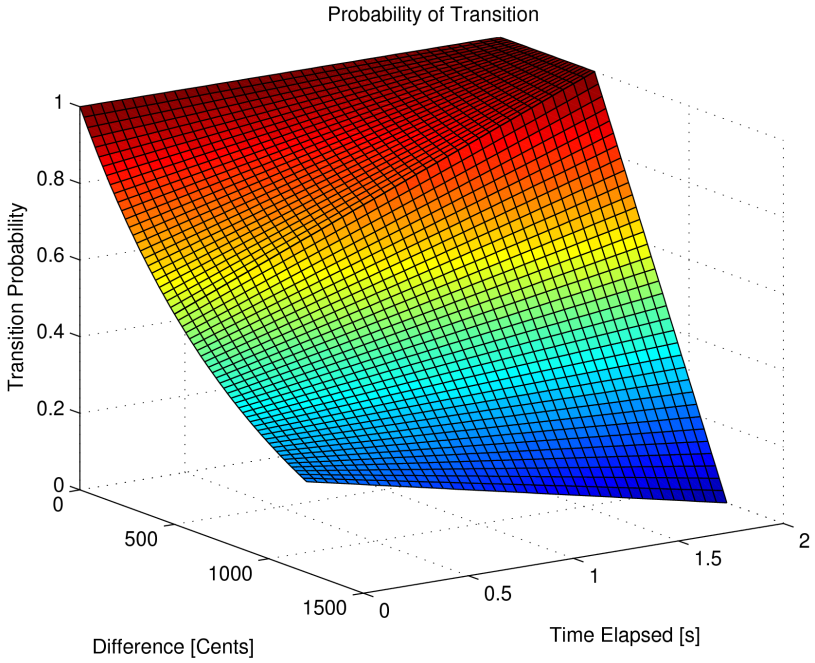


Figure 4.5: Transition probability function depending on difference in cents and on time

PDA	VE [%]	UE [%]	VE+UE [%]	GEH [%]	GEL [%]	GEH10 [%]	GEL10 [%]	DE [%]	HE [%]
ACF freq	44,4	23,5	31,6	1,2	0,1	1,5	0,18	0,4	0,06
DFE	26,6	15,5	20,4	8,4	4,2	16,5	8,9	0,2	1,3
MNBFCv1	22	12,7	16,3	0,4	21,2	1,5	22,1	0,06	19,5
MNBFCv2	22	10,7	15	0,4	1,1	1,8	2,3	0,03	0,8
MNBFCv3	18,5	13,3	15,3	0,5	1,2	1,9	2,6	0,05	0,8
MNBFCv4	15,6	16,3	16	0,6	1,3	2	2,8	0,05	0,9
MNBFCv5	22	10,7	15	0,7	1,7	2,1	2,9	0,14	1

Table 4.6: Channel 0 overall results

the first variant with the conversion difference to cents. MNBFCv3 is almost the same as MNBFCv2, but has the V/UV threshold set to 0.45, whereas MNBFCv4 has this threshold value set to 0.4. The final MNBFCv5 involves adding the temporal domain to transition probability function with the time forgetting threshold set to 2 seconds. Table 4.7 presents a comparison of the precision over five distinct frequency bands. To compare our method with other widely used methods, we added the results for autocorrelation in the frequency domain (ACF freq, a very good method for tracking singing) and the Direct Frequency Estimation method (DFE) [103], which is currently used for evaluating Parkinson's disease at FEE CTU Prague.

The results show that MNFBC is better than DFE in V/UV detection and also in precision. The VE+UE parameter is the best for MNBFCv2, but we can achieve the

PDA	57-88 [Hz]	88-141 [Hz]	141-225 [Hz]	225-353 [Hz]	353-565 [Hz]
ACF freq	92,7	5,2	0,6	1,1	17,7
DFE	26,4	12,1	12,3	13,0	53,4
MNBFCv1	2	1,9	28,3	49,7	73,7
MNBFCv2	1,1	0,7	1,7	2,2	32,1
MNBFCv3	1,7	0,9	2	2,3	33
MNBFCv4	2,3	1	2,3	2,7	34,4
MNBFCv5	2,2	1,6	2,8	2,9	32,1

Table 4.7: Gross errors (GEL+GEH [%]) in 2/3 octave frequency bands on Channel 0

best VE ratio for MNBFCv4 (but with a worse UE rate). The choice of variant depends on aimed application - whether we need to minimize voiced errors or unvoiced errors. For example, in the case of the planned punctuation detector we are trying to minimize the unvoiced error rate in order to obtain only confident F0 estimates. The results also show a big increase in precision with frequency difference (MNBFCv1) to cent conversion (MNBFCv2). Progress can be seen mainly in GEL and in the halving error rate. Table 4.7 shows that most errors for MNBFCv1 occur in the highest band, where the differences from the current frequency are much greater in Hz units than for lower bands. Thus these transitions are evaluated with very low probability, leading to these errors. The table also shows that the ACF method can provide the best results for the for highest frequency band, but is very poor in the lowest band. MNBFCv5 with temporal forgetting could probably not show its strength on the reference corpus because of lack of suprasegmental prosodic phrases (the corpus consists mainly of isolated words). But in comparison with MNBFCv2 however, there is only a slightly higher GEH rate. G lobally MNFBC with the addition of the Viterbi traceback procedure outperforms DFE on close talk channel 0. Note that it has much lower GEH and GEL even for lower VE. T his is not easy to achieve for PDA, because lower VE means that more uncertain segments (which other PDAs with higher VE have considered as unvoiced) pass to computation of the precision of F0 detection (gross errors).

Other results not presented in this paper have also shown a noticeable decrease in precision on channel 1 with the algorithm presented here. To get good results even in noisy environments, higher noise robustness is needed. This could be accomplished by adding a pre-processing stage with noise reduction (not implemented yet).

Conclusion

We have described a pitch-detection algorithm purely based on merged normalized forward-backward correlation (MNFBC) with an advanced Viterbi post-processing procedure for

finding the most probable pitch track. The optimal range of the voicing threshold was found for the MNFBC function. The results confirm that computing the transition probabilities with the pitch difference measured in semitones significantly improves the gross error rates (especially frequency halving) over the case of direct difference of frequencies. We have also tried to extend the transition probability function with a temporal dimension. This enhancement should lead to fewer errors occurring on the edges of prosodic pauses, but this has not been proven in experiments performed on a pitch reference database. This could be due to the very limited presence of supra-segmental prosodic pauses in the corpus. More experiments on suitable utterances need to be performed in order to evaluate this hypothesis.

Chapter 5

Prosodic databases and used material

5.1 Discussion about suitable prosodic database

After a few initial (and mostly unsuccessful) experiments regarding the Czech prosody it was realized that presented research is extensively dependent on suitable prosodically annotated Czech data.

As the term 'prosodically annotated' database might be quite ambiguous, in our ideal point of view it might be the one that contains following labels: individual phone time segments (or at least vowel lengths), syllable intervals, word segments, perceived delimitation of words into stress-groups (crucial information, optimally with the information how the stress-group was formed if it is multi-word stress-group), phrase delimitation with its perceived modality including the stress-group that bears the melodeme (and possibly with a note about the type of the question for interrogative phrases), breath-groups scope and finally emotional state (or attitude) of the speaker. As for acoustic signal description or required properties the audio should be recorded with focus on constant distance of the speaker from the microphone to obtain comparable intensity levels at least during the phrase. Database should contain some kind of manually verified pitch information in terms of: pitch-marks (specific points defined during vocal cords cycles) in signal, F0 values in voiced regions verified by listening (with notes regarding the creaky/breathy voice if present), stylized pitch contour that does not by any means compress the information perceived by the listener, annotated pitch accents that match the language set and speech style (this information should be as an addition only to the one or more preceded). Also, a very useful information included would be syntactic and semantic dependency structures for given phrase/sentence for deeper analysis of more specific phenomena that might

occur in given language.

As you can see, the amount of information needed for useful and valuable prosodic database is quite large and very few of it is currently reachable by pure machine processing. It rather involves close participating of language specialist (preferably a phonetician) who can (ideally) reliably annotate the data in reasonably short time. The inter- and intra-labeler agreement of prosodic properties can be also quite troublesome fact, especially in the field of stress-groups (as referred e.g. by [118] and from our personal experience).

For phonetic kind of research, where we typically operate on tens, maximally on hundreds of samples to run various statistical tests for rejecting/not-rejecting investigated hypothesis, this spreading feature set is still feasible. But for machine learning approaches (including speech recognition) where we typically operate on ten-thousands of samples (or prefferably more), this kind of magnitude is very challenging to manually process. This is why our research tries to balance on the edge between both approaches with our currently very limited dataset sizes.

As one can see, the stress-group segmentation information in consequence with sentence modality information and rich content in terms of modality is crucial for the purpose of our research. This is why common commercial corpora as SPEECON, SPEECHDAT and even pure Broadcast News (without any interviews) are without any further labeling simply insufficient for the kind of research it is dealt with in this thesis.

5.2 Used material for stress-group issues

This section brings an overview of used corpora in the experiments described in the following two chapters.

- Czech SPEECON (phonetically rich read sentences subset):
  The corpus consists of 10,062 utterances (491 speakers, 72,042 stress-groups, 181,640 syllables). Corpus is aimed for stress-group detection task and obtaining statistics for feature normalizations.

  Author later filtered special 'clitics-free' subset from the initial set, with magnitude of 630 utterances (about 9,000 syllables). Out of those were extracted only those that did not contain any monosyllabic word and set of Out of this 'clitics-free' subset, 189 utterances (153 speakers, 1150 stress-groups, 3344 syllables) were manually

verified by the author in terms of syllable nuclei center timestamps together with stress-group delimitation.

- Corpus from diploma thesis of Helena Spilková [76] (read utterances):
  Corpus consisted originally of 152 utterances with ambiguous possibility of segmentation into stress-groups, the list of utterance pairs can be found in appendix C. The set was filtered so that it only contains utterances where speaker's aim is in accordance with author's perception (130 out of 152 utterances meet this condition). Corpus already contained phone-level segmentation information within Praat TextGrids. Filtered corpus magnitude is 130 utterances (4 speakers, 427 stress groups, 1269 syllables). Corpus is suitable for stress-group detection task and was also used as a proof of concept experiment dataset (section 8.3).

- Corpus from diploma thesis of Eliška Churáňová [119]:
  Read broadcast news by professional speakers. Available part for the presented research are manually verified utterances in terms of stress-group segmentation and phoneme alignment done by experienced phonetician. Labeling information is contained within Praat TextGrids. Available data consist of 364 utterances in total read by 6 speakers (9293 syllables). Data are suitable for stress-group segmentation tasks and due to the high speech rate they can be considered as a transition from simple read utterances (SPEECON style) to totally spontaneous speaking style.

5.3 Used material for phrase modality tasks

This section brings an overview of used corpora in the experiments described in this chapter.

- Author's manually selected collection of Czech audiobooks phrases:
  Publicly available data, read by leading Czech actors, MP3 format (studio ambient), guarantee of rich prosodic material. Parts of 4 audiobooks read by 4 different actors (3M+1W) were selected and manually annotated. With consideration of future automatic processing of the corpus, whole sentence intonations were taken as pitch patterns (between neighboring punctuation marks).

Chapter 6

Sentence division into stress-groups

The task of sentence division into stress-groups can be also understood as finding of lexically stressed syllables. In fixed-stress languages the stress-groups segmentation of utterance is very close to the word segmentation of sentence, except for the fact that several words might join into single stress-group. This means that the specific word composition of sentence is prosodically and acoustically hidden and the closest observable (or more correctly perceivable) realization is expressed exactly by the stress-group segmentation. A successful mastering of the task may thus be perceived as the essential condition for sentence segmentation into words or for word boundary detection given the acoustic signal.

6.1 State of the art in Czech

In this section we bring an overview of works related to the problem of lexical stress or stress-group detection in Czech language. The initial qualitative (and quantitative in terms of statistical tests) overview is already covered in the section 3.4.2. Here, an existing quantitative research typically related to machine learning applications is presented instead.

Experiment question in paper [120] was: "How can Czech accent be inferred from prosodic parameters only (without lexical information) using ANN?"

Material consisted of read newspaper article with 1100 syllables by one semi-professional male speaker, 30% of them accented (after manual accent labeling). Used features were: raw F0 over whole syllable [Hz], normalized syllable nuclei duration, normalized average intensity of nucleus (normalization used to suppress intrinsic factors of different nuclei vowels). Used classifier was feed-forward MLP ANN with 2/3 train set, 1/3 test set Tested configurations were 3-syllable context driving 9 inputs into 18 hidden neurons and extended 5-syllable context windows driving 15 input features into 30 hidden neurons.

Each context was also tested with various position of actual syllable: forward, centered and backward window (with inter-steps for 5-syllable context). Results showed that one-sided context is less optimal then centered contexts, which confirms contextual hypothesis of Czech accent. The 5-syllable context windows showed slightly better results, although the statistical improvement of result over 3-syllable window was not proved. Despite the fail of keeping F0 values in [Hz] units and not converting them to semitones, the trained ("speaker-adapted") network was able to attain an 80% agreement with human listener. Work also contributes thorough error analysis section.

Paper [121] follows the pilot study [120]. Primary question to be answered was, whether or not perceived accents in Czech have objective existence. It is answered in positive way in the introduction part by giving indirect evidence of accent presence in Czech. Secondary, "what are respective roles of used features (pitch, duration, intensity) in accent prediction?" The features are compared individually, in pairs and all together. Dataset comprised of an informative text read by one male semi-professional speaker containing 970 syllables. Accent labels were manually assigned to syllables according to the authors perception, which leaded to the fact, that 32% of material were accented syllables. Compared to used features in the previous pilot study, this time authors used two-fold normalization for durational feature. Newly, authors operate with 3 intonation values per syllables (in 20,50 and 80 % of its duration). Unfortunately, there is still missing conversion from [Hz] to musical units [e.g. semitones] which correspond to human frequency perception. Regarding the ANN network configuration, again, feedforward MLP ANN was trained with back-propagation algorithm. 3-syllable context window with trained syllable in the center was used leading to 15 inputs leading into 1 hidden layer with 30 neurons. Results were, that F0 alone is a good predictor of stress (prediction score around 78%). When combined with other features the results get slightly better. For all three tested prosodic parameters involved, the prediction score was 80%. The accent predictions for each syllable were made independently of previously assigned accent values. Study presents thorough error analysis as its predecessor suggesting the improvement if the network could do rhythmical predictions. Conclusions fro the study were, that although improved normalization and more detailed F0 contour information provided to the classifier, the result did not outperform those from their previous pilot study. Besides, study showed relative importance of the individual investigated prosodic parameters in accent characterization. Redoubtable critical contexts as adjacent accents and anacruses did not generate worse than average results. Future work suggested comparison with other languages with "weak" lexical stress.

More realistic name of the paper [122] might be: "Variability of normalized intensity in read Czech stress units and across them" as the author himself claims. Paper brings an acoustic description of macro-intensity patterns of stress units in read Czech using normalized intensity of syllable nuclei. Material consisted of 3 male speakers, each read approximately 900 syllables, 2700 syllables in total. Speakers were trained to keep the distance from the microphone during whole recording session. Experimental setup was following: representative intensity feature was extracted as intensity mean over the vowel segment. Intrinsic factors of particular vowel class intensity differences were excluded by normalization of raw mean intensity values and anacrusis type of stress-groups were excluded (4.4% of original sample). Author provides a nice graphs of individual average speakers normalized intensity contours within stress-groups of various lengths. The conclusions was, that there is very good match of all significant results compared to JND of intensity in syllable nuclei [123]. Normalized intensity values show gradual macroprosodic decrease over inter-pause groups, followed typically by significant intensity reset. Local intensity drops occur between the last two syllables of the stress unit, with a major intensity drop before the pause. Syllables bearing perceived accents do not show intensity peaks, while initial intensity rise in longer units seems to be speaker dependent. It is mentioned, that this paper does not take into account relations and interactions of intensity with neither F0 (intonation contours) and duration (final lengthening), nor spectral balance. Nevertheless, in conclusions authors claim that their data support the hypothesis of a trade-off relation between pre-boundary lengthening and dynamic decreases.

Work [124] deals with detection of emphasized words (focus) in Czech. Although it is a task not same as accent realizations coming from potential lexical stress, it is described in this section, because the term focus can be often misinterpreted with term stress or accent and its difference in terms of acoustic realization is needed to be emphasized.

Applied corpus consisted 180 sentences recorded specially for this experiment. Three speakers were asked to put emphasis on arbitrary word in sentence. The task was therefore to detect sentential stress as realization of speaker's conscious focus on particular word (rather than lexical stress detection this thesis is dealing with). Authors claim that initial syllables of focused words in Czech are generally characterized mostly by intensity increase, plus there is some increase in duration and minor increase in pitch. Authors then detect emphasized word in the utterance by simply summing the normalized contours of all of these qualities and the syllable with highest peak is then considered as the beginning of the emphasized word in the utterance. Their system achieves overall score of 91% in the task of emphasized word detection. The most significant feature alone was found to be relative word prolongation with 86% score.

The results clearly indicate the difference of acoustic realization between accent as realized lexical stress and focus (typically on syllable level) as an emphasized word in sentence.

6.2 Czech Text-to-Foot converter

For the purposes of automatic labeling of Czech sentences into stress-groups for automatic training of stress-group models on higher amount of data, a prosodic module covering the most unequivocal Czech clitic absorption rules was implemented [125]. Module automatically converts given sentence into stress-group (foot) units with count of syllables in each of risen foot. It is supposed that such a tool can be really useful for various prosody based speech recognition systems for Czech in the future. A complete information on Text-to-Foot (TTF) converter for Czech including theoretical background, implementation details and tool evaluation follows.

6.2.1 Relationship between Czech words and stress-groups

As already stated in 3.4.2, the words often tend to keep their autonomy and create their own feet in Czech continuous speech. Sometimes the word autonomy is canceled and the word creates one foot together with the other neighboring word. The general tendency is that multi-syllable words create their own feet, the only exceptions are single syllabic words. For them, these realizations into foot division can occur:

a) Single syllable word joins to its predecessor and loses the prominence. This tendency is very strong and such words are generally marked as enclitics. For some of nouns this tendency is almost a rule. Another category that can follow this tendency are adverbs ("snad", "dnes", ...), but particular realization depends on the context and order of the words in the sentence.

b) Single syllable word joins to its successor and the prominence of the successor is converted onto the single-syllabic word, that creates the beginning of the newly risen foot. This is typical for one-syllabic prepositions (proclitics) and the tendency is again very strong.

c) Single syllable word is followed by another one syllabic word. In this case both words join together and two-syllabic foot is created with the prominence on its first syllable. This situation can chain with another single-syllabic word.

d) Single syllable word keeps its prominence and creates its own foot.

Some of the conjunctions (a, i, že, ...) create special group of one syllable words called "předrážky". They join to the following word, but do not have prominence. This is the

very special case contrary to the basic rule of the prominence on the first syllable of the foot. Besides, there are theories that there is kind of minor prominence on odd syllables in multi syllable words in Czech (on the third, fifth, ...). But according to some researches, the minor prominence is not really present in spontaneous speech at all [74].

6.2.2 Text-to-Foot converter implementation

Text-to-Foot Converter was implemented in multi-platform Perl scripting language with only default modules used and tries to follow the phonetic and grammatical theory. Module gets exactly one Czech sentence on the input (with or without punctuation marks, sentence leading capital letter is also not mandatory to accomplish the generality of the module) and using a defined set of transforms it iteratively converts the text into the the foot units. The task can be also understood as the mapping of the word sequence to the foot sequence. T he performed text transformation is always the deletion of leading or trailing space between particular words (depending if they are enclitics or proclitics).

0. prephase: non-syllabic prepositions ('v','k','s','z') are absorbed to the following noun
1. phase: transformation of enclitics (especially pronouns and special forms of verb 'být'), with following order of absorption [126]:
   (a) category 1: jsem, jsme, ... + bych, bys, ...
   (b) category 2: si, se
   (c) category 3: mi, mně, ti, mu, ...
   (d) category 4: mě, tě, ho, ...
   (e) category 5: mnou, něj, něm, ... + prý, však
2. phase: transformation of proclitics (especially prepositions - u, o, ze, ke, ve, na, ...)

Note that the order of phase 1 transformations for the enclitics is mandatory, it is needed to happen from the leftmost enclitic one by one. If the order is not kept, there is a risk that some enclitic will remain unjoined with the previous word as it cannot then be matched by its pattern (because some of other enclitic could already join to it from the right side).

First set of rules was created for the patterns located in the middle of the sentence (delimited by space sign on both sides). Another option is the case when proclitics stay in the beginning of the sentence, in this case we need to write explicit rules for them. Similar situation is for enclitics located at the end of the intonation units and succeeded by the any type of punctuation mark. All these rules had to be added too. One-syllable

adverbs and connectors does not follow any hard rules and generally can act as both types of clitics. That is why their processing was not implemented in the current version of the module.

With such defined rules an example of processing the input sentence "Zeptal bych se ti v pátek." could be following:

1. prephase: string ' v ' found $\Longrightarrow$ "Zeptal bych se ti vpátek."
2. category 1: string ' bych ' found $\Longrightarrow$ "Zeptalbych se ti vpátek."
3. category 2: string ' se ' found $\Longrightarrow$ "Zeptalbychse ti vpátek."
4. category 3: string ' ti ' found $\Longrightarrow$ "Zeptalbychseti vpátek."
5. category 4: nothing found
6. category 5: nothing found
7. no proclitic found
8. final output: "Zeptalbychseti vpátek."

Final sentence on the output of the converter is therefore "Zeptalbychseti vpátek." which denotes that whole sentence was converted into 2 feet only: first foot with length of five syllables (label 'FOOT5') followed by two-syllable foot ('FOOT2.'). In this case, any other order of rules for categories 1-3 would lead into wrong concatenation of words and final sentence would have more than two feet.

6.2.3 Used data set

The main database for whole foot-detection system is manually selected part of the standard Czech SPEECON speech database. The Czech SPEECON database comprises the recordings of 550 adult and 50 child Czech non-professional speakers. We used the cleaned subset of the database (but about 90% of the data remained). Moreover, we took only utterances containing whole sentences from this subset. Thus, our initial data set was composed of 12,345 sentences. But there was some more filtering needed: especially double vowels (diftongs) are problematic in Czech. There are two natural diftongs in Czech 'au' and 'ou', both are mostly pronounced as one syllable (there are some exceptions but they are handled by our system well). The other diftongs were rather filtered out due to their generally ambiguous syllabification which is difficult to decide (sometimes even for native speaker, e.g. in words 'neutron', 'pneumatika'). We did the following filtering of testing material at the input stage:

- 'eu' sentences omitted - (2,079 sentences filtered out)
- 'eo' sentences omitted (117 filtered)
- 'ao' sentences omitted (32 filtered)
- some other filtering was done (sentences beginning with 'Ó ...', 'Jó ...')

After this filtering we obtained final subset of 10,022 sentences which were then processed to obtain their foot division and also some prosodic features used later for training of HMM models in the experiment described in section 6.4.

6.2.4 Results after application on data

Having all input sentences transformed using the TTF converter, we can do some rough statistic comparison of foot length occurrence within our material with existing statistics of the Czech foot lengths [74]. The referred values in the table 6.1 are not exactly overtaken, because exact references originate from material with distinguished direct and indirect speech (50,000 feet in total). Our corpus is something between both types of the speech, that is why average-like values of these references were taken as final references values. Another issue is "předrážka" phenomenon that originally created its own category in the overtaken statistics. To be able to compare the whole set of categories we joined "předrážka" group with one-syllable feet (FOOT1). We have labeled 72,353 feet after the text transformation in total in our data set.

FOOT type	**Our material [%]**	**Reference [%]**
FOOT1	21.7	15
FOOT2	36.3	39
FOOT3	26.8	32
FOOT4	11.2	13
FOOT5	3.3	3.2
FOOT6	0.5	0.5
FOOT7	0.05	0.05
FOOT8	0.0	0.01

Table 6.1: Occurrences of specific foot lengths derived from tested dataset using Text-to-FOOT converter in comparison with reference distributions

6.2.5 Results discussion

Conclusions from the foot occurrence comparison in table 6.1 is that our transformed data show very similar proportion as reference text statistics. The only except is the FOOT1 (one-syllable foot) category, where there are still about 5% more occurrences than in the referred statistics. These extra one syllable words are most probably the adverbs that should be joined to one of their neighboring feet and thus create two (or more) syllable feet. These cases cannot be probably covered by hard-driven rules so-far implemented in our foot converter module, but do need some higher level of knowledge about the text (context, neighboring words as part of speech etc.). This is in accordance to the observation that we obtained less two-syllable feet (FOOT2) and even less three-syllable feet (FOOT3) than reference material. Groups FOOT4-FOOT8 show very similar occurrence ratios in both corpora.

We also measured number of occurrences of single-syllabic words in our original data set (but after absorption of non-syllabic prepositions) and a number 32,409 was obtained. With comparison of 15,724 one-syllabic feet obtained after the text-to-foot conversion we achieved "footing factor" of 2.06. It would be interesting to compare this value with some phonetic statistics if they existed. But if we aim to have 15% of single-syllabic foot (out of all the feet) after the conversion, this factor needs to be around 3.

6.3 Used features, extraction and normalization

According to Czech lexical stress being marked as weak and there are assumptions about trade-offs in realized acoustic qualities, it seems to be highly important to extract and normalize the acoustic features in the best available manner.

6.3.1 F0 extraction methodology

6.3.2 Choice of PDA

The choice of PDA method Regardless any interesting results presented in 4.2, even after years of its presentation, the Praat [112] autocorrelation pitch detection method seems to be preferred choice in the field of phonetics and speech processing. In [127] it is claimed that 12 out of 19 forensic phonetics used Praat software for fundamental frequency analysis. It also seems that Praat is the choice for most of papers published in recent years considering deriving the prosodic features. It is understandable because of the possibility of fine tuning the parameters for the methods, quite easy batch applicability of methods on huge amounts of data via own scripting language and also for simple reproducibility

of experiments for interested researchers. This is why raw F0 values presented in the following research were obtained from audio signal using Praat autocorrelation method.

6.3.3 F0 extraction approach

There are various possibilities of pitch-related feature extractions.

- continuous contours or discrete tonal approach:
  Both options were already discussed in 3.6 based mainly on [7]. In summary, there are more valid methods:
  - create continuous contour from raw non-continuous PDA output (and probably suitably filter the microintonation)
  - extract pitch descriptors only at the one right point per syllable/vowel
  - extract more pitch descriptors within syllable/vowel (e.g. at 3 distinct points, but 3 values per vowel did not improve the accuracy of experiment in [121])
- possibilities for discrete pitch descriptors (what to extract):
  - just 1 direct pitch value at given point
  - 1 direct point + 1st derivative as pitch velocity around this point
  - 1 direct point + 1st derivative as pitch velocity around this point + 2nd derivation around this point as concave/convex property of pitch curve

The extraction method chosen in presented research is based on the tonal approach ([7] and [93]) claiming, that audible pitch changes occurring in the syllable nuclei are represented reliably. As used automatic ASR force-alignment or any manual labeled data do have timestamps either directly for syllable nuclei center or vowel boundaries, the tonal approach can be beneficially used and thus extracted F0 values correspond to the syllable nuclei centers.

6.3.4 F0 normalization methodology

We can divide the normalization approaches into following situations:

1. **Unknown speaker**:
   How much data do we need to gather reasonably good statistics so the speaker becomes "known"? This studied Volín in [128] and he suggests that at least 20 seconds of "speech only signal" should be analyzed for Czech so that examined descriptors are stabilized enough and valid F0 statistics can be obtained. The

reliable and pretty safe amount seems to be on the other hand a magnitude of 10 utterances. If the there is not enough of data available, the most reasonable way seems to be an estimation of utterance average pitch value and obtaining semitone pitch values related to this average.

2. **Known speaker**:
   In this case it is expected that speaker's F0 statistics are already available or there are enough data to obtain them.

   In the initial "1st-level" F0 normalization, the aim is to suppress differences between speakers. Having long-time distributional measures covering an overall height of the speaker's voice (pitch level) and the range of frequencies covered by the speaker (pitch span), there are two options how to process pitch data:

   (a) speaker's pitch-level normalization
   Usual approach for obtaining speaker's pitch level is to use kind of central tendency, which can be covered by mean or median. Our research is rather based on using 'baseline' F0 of overall height of speaker's voice as the normalizing pitch level. Baseline F0 is seventh percentile, proved to be a stable characteristic across various speaking styles and/or recording conditions [129]. All speakers' pitch estimates are related to this baseline pitch level in the following steps: Firstly, all pitch values already in semitones related to each utterance mean value are converted to new semitone values related to 100 Hz. Having all the pitch values related to some fix frequency, the histograms like statistics can be obtained for each speaker and his 7th percentile is derived (actually using R-script) as speaker's baseline F0. Followingly, all the pitch values are recomputed to be related to computed speakers' baseline F0. Described normalization keeps pitch data in pure musical unit semitone domain and I believe its application can be only beneficial to any pitch related studies.

   (b) speaker's pitch span/range normalization
   There are more options how pitch span can be defined: by standard deviation (needs to be computed on musical interval scale due to the asymmetricity of differences in [Hz] units), by a variation range (the difference between maximum and minimum; can be computed on both musical and Hz scales), by the 80-percentile range (the difference between the 90th and 10th percentile) and finally by the quartile range (the difference between the 25th and 75th percentile). Last two parameters display more stability than the maximum-minimum range, since certain portions of extreme values are cut off [130].

   Also, percentile/quartile ranges can be computed on both musical [ST] and

frequency [Hz] scales as we are interested only in the order of values in this case. Just note, that in the case of reporting the percentile/quartile based pitch span statistics in [Hz] units, the whole interval is needed to be reported, not just the difference itself.

The pitch normalization options for known speaker lead to the hypothesis, whether does exist some global tendency in Czech, when the speaker realize the stress-foot delimiting by some absolute (musical) contours regardless to their overall utterance intonation spans or does the pitch stress-foot delimiting differences corresponds relatively to speaker overall pitch span. The answer to this question might be realized by 2 different sets of pitch features, that might contain:

(a) pure 'musical' pitch features only related to baseline F0 pitch of each speaker,

(b) all the speaker's pitch values scaled within his/her speaking range (with "destroyed" musical unit domain).

As PDA algorithm with non-filtered output might still contain some octave errors (just to note, those octave errors might come from real physical halve-octave vocal chord mode in phone transitions or from creaky voice). Here I describe the currently used enhanced approach for F0 baseline estimation and pitch span normalization:

I claim that prior to the (a) (speaker pitch-level normalization) and (b) (pitch span normalization), right after obtaining speaker's statistics that might be biased by the presence of octave errors, there is a simple way how to lower their impact on F0 statistics precision.

Firstly, from the speaker's F0 statistics (semitones converted, even with octave errors) a median F0 value is taken. Secondly, we continue with filtering out F0 content that lies outside the $< -10; 10 >$ ST interval around median as they very likely present unwanted outliers. From manually checked histograms centered around F0 median value for each speaker in our SPEECON subset it seems that -10 and -10 ST are the valid and quite universal borders that delimit valid F0 values from octave errors. For comparison and for an acceptance of the idea by common sense, the given 20 ST 'valid' pitch range is more than one and half octave (18 ST), which is very close to the total singing range of untrained singer. It is also clear that during the speech the total singing range is rarely used completely and speech pitch span is rather compressed compared to the singing one. To be sure about correctness of this procedure on real data, one may want to know, to which percentiles those values -10 and +10 ST correspond for individual speakers in data set. This can be achieved using empirical cumulative distribution function (e.g. 'ecdf()' function in R toolkit). Finally, after the filtering of unwanted outliers, we are ready to

obtain new values for both types of pitch feature normalizations (baseline F0 and chosen representation of pitch span).

The procedure can be illustrated by the figure 6.1.

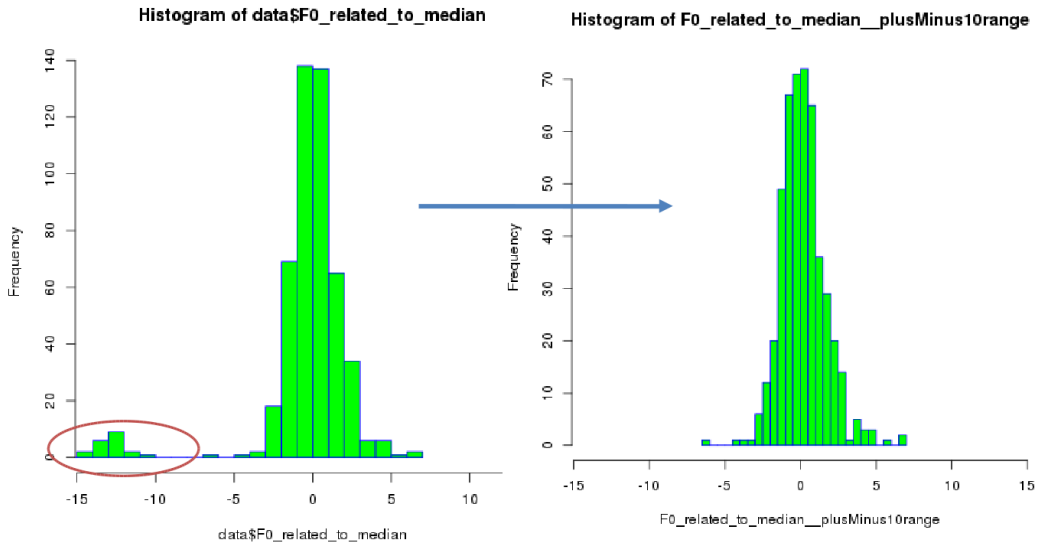


Figure 6.1: Example of octave errors filtering in F0 histogram for obtaining better F0 statistics for given speaker

6.3.5 F0 extraction post-processing

Octave errors correction method

The initial state is to have discrete F0 features related to speakers baseline F0. This allows to detect weird F0 estimates using "absolute" musical distance from the F0 baseline. Empirically stated thresholds for are:

- -8 ST as low threshold,
- +11 ST as high threshold

Continuously, we check the differences and judging what is still relevant pitch jump and what might be potential PDA artifact. Being rather conservative, the empirically stated symmetric value is 8 semitones for validity of F0 jump. Lower symmetric threshold brings the false positive errors with an effort to repair even valid F0 jumps. The "suspicious" jump is the one with the immediate return, the threshold here was set to 7 semitones. Our octave error fixing algorithm is able to repair even two halving errors in a row. A real utterance with such artifacts is depicted in the figure appendix B.1.

6.3.6 Intrinsic F0

One might come with a question, if there exists something like intrinsic pitch for individual vowel identity as an analogy to intrinsic intensity of individual vowel identity. We initially wanted to measure those tendencies ourselves, but such experiment requires various conditions to be met. Optimally, we should take a long enough fixed C/V context in which will only vary the center with various vowel identities. This condition is for sure very difficult to meet and the closest feasible approximation might be to get at least the same vocalic context within stress groups of same lengths. But even this approximation cannot in our honest opinion lead to satisfying results as the pitch contours realizations will vary up to the extent that will ruin the intrinsic pitch differences that are expected to be in lower magnitude. The study of this phenomenon thus needs a fully controlled experiment with carefully planned content of utterances. For now, we will settle for the information given in the literature. According to [128] and [63] the phenomenon exists, but is rather more complex: in the speech production end of chain there figures 'intrinsic F0'. Thanks to it, the higher vowels like [i] and [u] reach on average higher F0 values than lower vowels under identical conditions. But these differences are compensated by human perception mechanism and this phenomenon on the consuming end of transmission chain is called 'intrinsic pitch' being antagonist for intrinsic F0. High vowels are than actually heard on lower pitch than low vowels of the same F0 [6].

Presented experiments do not deal more with intrinsic F0 factors, although they do exist.

6.3.7 Intensity

In real analog physical world, the energy of wave signal can be converted to intensity taking into account the time dimension and an area (intensity being the power emitted into unit area). In acoustics, by the term "intensity" it is often meant sound intensity level (SIL), which is by normal condition almost equal to sound pressure level (SPL) in decibel [dB] (see section 2.1.4 for details). SPL level is on the other hand measured acoustic pressure related to the pressure of 2.10^{-5} Pascal, which is the human normative auditory threshold for a 1000-Hz sine wave.

In the digital world the signal went through the sampling process being represented by zeroes and ones in typically 2 byte per sample amplitude resolution. The way how Praat [112] software (which is used to extract the intensity features from audio signals in presented research) operates is following: 16-bit digital signal value range in 2B signed

integer representation is between -32768 and 32767 (these extremes represent maximum positive and negative sampled amplitude). Sample values are normalized to be in the "float" range between -1 and 1 by dividing the sample values by 32768. Next, Praat considers obtained values to be directly air pressure values in units of Pascal to compute sound intensity in SPL [dB] scale according to equation 2.4. Therefore, the maximum value of intensity SPL observable in Praat output of intensity object is $20log(5.10^4) = 93.98$ dB. Any higher value would cause audio signal clipping. Also, this means that values of intensity SPL obtained are not true representatives of original signal sound pressure levels unless there was taken simultaneously a reference measurement for signal of known SPL intensity (and no such referenced speech database is available for presented research). Fortunately, regardless of possible shift in absolute intensity values reported by Praat, all the differences (e.g. signal to noise) in values are still valid thanks to the nature of decibel unit.

Intensity normalization

- "blind" intensity normalization:
  In this case, we do not have particular vowel identity information and we do not know the vowel positions in signal. We can still normalize the signal using following procedure. In our setup we use this kind of normalization as one of methods for detection of illegal nuclei center timestamp obtained from force-alignment. One possible method for automatic energy normalization is estimating and applying a SNR-like estimate.

  1. Overall average intensity is computed from the whole utterance.
  2. We divide the intensity values into two groups according to the relation to computed overall average and we compute new averages for both groups so we get high average (AvgHigh) and low average (AvgLow). The difference of those two "biased" averages can be interpreted as first estimate of signal-to-noise ratio of the signal (utterance). It should work reasonably well for the signals, where speech is at least several dB higher than background noise and considering the noise part (including pauses) is roughly same in length as the speech signal. In that case the AvgLow should match the noise level and AvgHigh should match the speech level.
  3. Having SNR-like estimate we are than able to relate the actual signal intensity Int(t) and obtain its normalized value $Int_{norm}(t)$ using following formula:

$$Int_{norm}(t) = \frac{Int(t) - AvgHigh}{AvgHigh - AvgLow} \quad [-]$$

Using this kind of normalization, we might expect following values with given interpretations:

- $Int_{norm}(t) = 0$ value, exactly same intensity as high average.
- $Int_{norm}(t) = -1$ value, exactly same intensity as low average.
  This should not happen in reality and if, it might be a notification about false nuclei timestamps (in background noise part of audio signal)
- $Int_{norm}(t) = -0.5$ value, should be exactly at the whole utterance intensity average.

Advantages of presented normalization method is its easy implementation and reasonable robustness. Shortcomings are:

- utterances are NOT directly comparable, because final values depend on:
- ratio of speech and silence in the utterance (too much 'silence' against the 'signals', lowers naturally the global average and makes the SNR bigger for higher portions of silences in the signal
- utterance 'SNR' directly influence final values scale (low SNR utterances will have final values closer to each other than high SNR utterances)

- "partially informed":
  We do know the nuclei centers timestamps. While assuming the nuclei center times are correct, we can compute average intensity over all nuclei centers within given utterance and relate all the values to this average.

- "fully informed":
  Same as partially informed normalization, but vowel identity intrinsic intensity normalization is performed. Assumptions is, that we do have an information about the nuclei timestamp and particular vowel identity. We firstly relate all the intensities to the utterance average over all intensities in nuclei centers, then we apply the intrinsic normalization according to particular average intensity of given vowel in the corpus (see figure 6.2 for those values computed on whole SPEECON subset).

6.3.8 Syllable nuclei centers relative time distance ("relTime")

So called "relTime" (R_{Tn}) feature represents ratio of time distances between neighboring nuclei centers. One nuclei left-right context is needed, because the distance of actual

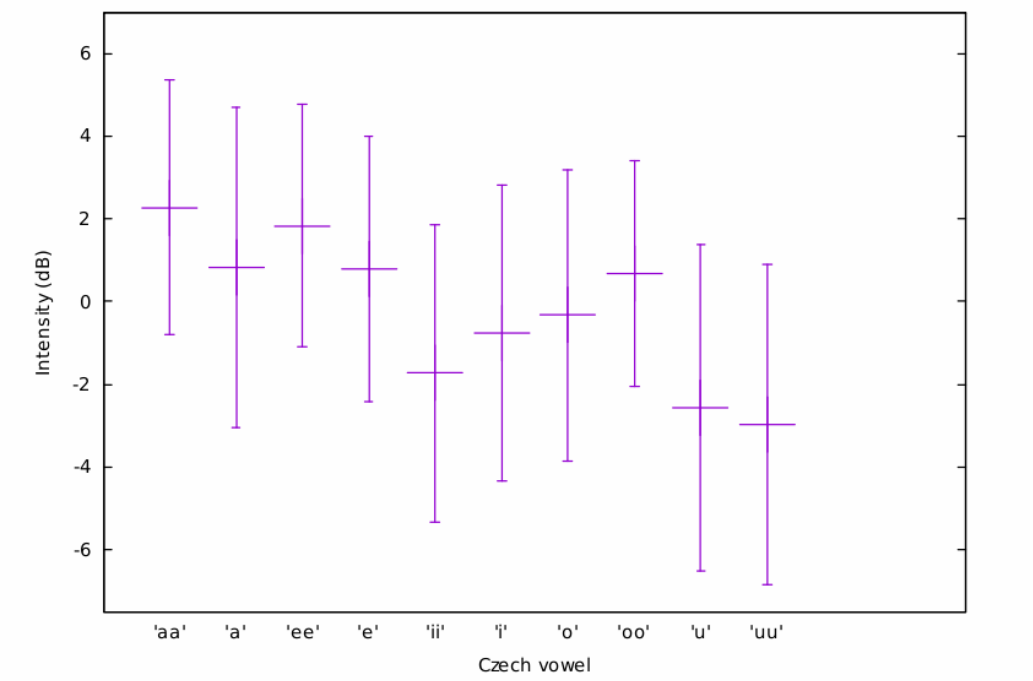


Figure 6.2: Mean intrinsic intensities of Czech vowels used for intensity feature normalization

versus previous nuclei and also following versus actual is computed (two distances between 3 points in time). The R_{Tn} is defined in equation 6.1.

$$R_{Tn} = \frac{t(n+1) - t(n)}{t(n) - t(n-1)} = \frac{\Delta t_{following}}{\Delta t_{preceding}} \tag{6.1}$$

By this single number R_{Tn} we hope we are able to capture to some extend a kind of temporal tendency between the three neighboring syllables. The value greater than one represents the "future" nuclei distance being greater than "previous" nuclei distance and should denote slowing down of speech. On the other hand, values smaller than one should indicate local speed-up of speech.

The very fist and very last nuclei in the utterance was defined to R_{Tn} value 1. Although we initially defined it to be 0, the redefinition better represents possible reality (that there might no time distance change) and secondly, it does not cue the machine learning classifiers to be "trapped" by those special values at the beginning or end of utterances.

We expect R_{Tn} feature not to be very representative without further normalization, but as we do not have accurate length of individual phones for the most of our current dataset, we are not able to easily normalize this feature. To correctly normalize the relative time distance feature, one also needs to know consonant/vocalic (CV) structure of syllable parts composing examined intervals (second half of previous syllable with the first half of the current and second half of current syllable together with the first part of following one). Firstly, neutral (from stress point of view) statistics for all C/V combinations should

be firstly taken. We are aware of the fact, that here lies a big space for improvement of the method, because it is assumed (and it was personally discussed with phoneticians) that some kind of fine temporal prominence is probably occurring in the scope of the first "stressed" Czech syllables, but it is not as strong as in other languages (while keeping the basic lengths division into short and long vowels to satisfy phonological condition and differentiate the meaning).

6.3.9 Spectral slope

Spectral slope features are extracted using Praat script as the energy sub-band differences. There were two variants of spectral slope features extracted according to [66]:

- fixed pivot with sub-bands: 0-500 Hz and 500-4000 Hz
- floating second formant pivot with sub-bands: 0-F2 Hz and F2-5500 Hz

Statistics for intrinsic spectral slope values of individual vowels were gathered and particular normalization was then applied.

6.3.10 Corpora processing and feature extraction system

Proposed feature extraction and normalization system has currently two modes of operation:

1. from force-alignment bootstrapped data
2. from fully manual labeled data

The diagram in the figure 6.3 presents a flow related to the first option, the input from ASR force-alignment.

In this case, we were facing with problem of nuclei center estimates precision. This is why we incorporated into the flow several control mechanisms, that allow to move falsely placed nuclei timestamps to the correct position using contours of pitch and intensity. The correction of nuclei center timestamp for pitch contour can be described followingly: We search for boundaries of "suitable" surroundings of F0 curve around tested nuclei center. "Suitable" surrounding is continuous and it does not exceed allowed F0 gradient threshold (an F0 estimation error or microintonation presence). If suitable surrounding is found in the range of $< -30; +30 >$ ms, everything is correct and we assume the nuclei timestamp is valid. If the suitable surrounding is not symmetric, we move the nuclei timestamp to the center of the suitable surrounding.

The diagram 6.3 can be briefly described as follows: Features are extracted and normalized using the methods described in section 6.3. Statistics of all extracted features can be gathered from the incoming data, or already pre-computed values on bigger corpus can be applied. The interesting option is generating the Praat textGrids files for each processed utterance, so it can be than manually verified and possibly corrected. If this is the case, the system has than option to import the input information from given textGrid files. The import mechanism is quite general and can thus be applied even for a new corpus with different textGrid configuration. Another feature of the system worth noting is generating of gnuplot source files for all processed features per each utterance with visual information about actual footing segmentation of the utterance together with the utterance transcript. The output of the system are configurable feature vectors with various target labels almost directly suitable for various machine learning processes.

6.4 Initial experiment

The initial experiment on Czech stress-group segmentation using pitch only information and concept of Hidden Markov Models was carried in [131]. The motivation was to examine, to which extent the melody of speech alone is in Czech a clue for stress-group delimitation using machine learning methods consisting of statistical model training.

6.4.1 Used material

A subset of read sentences by non-professional speakers from Czech SPEECON database consisting of over 10,000 utterances was used as the only material in the presented experiment (for details see section 5.2 for details).

A canonical stress-group segmentation was obtained from utterance text reference transcripts automatically by the Text-to-Foot conversion module only (module is in detail described in section 6.2), there was no further human validation due the volume of used data. Followingly, by application of Czech syllabification module on the stream of created stress-groups, the target labels consisting of string "FOOT" followed by the number N (denoting number syllables within stress-group) were acquired.

6.4.2 Acoustic features and their normalization

The only input features to whole system were fundamental frequencies converted to semitones related to each utterance average fundamental frequency (obtained from voiced parts). Not the whole pitch value series was used as the input, but only those values that

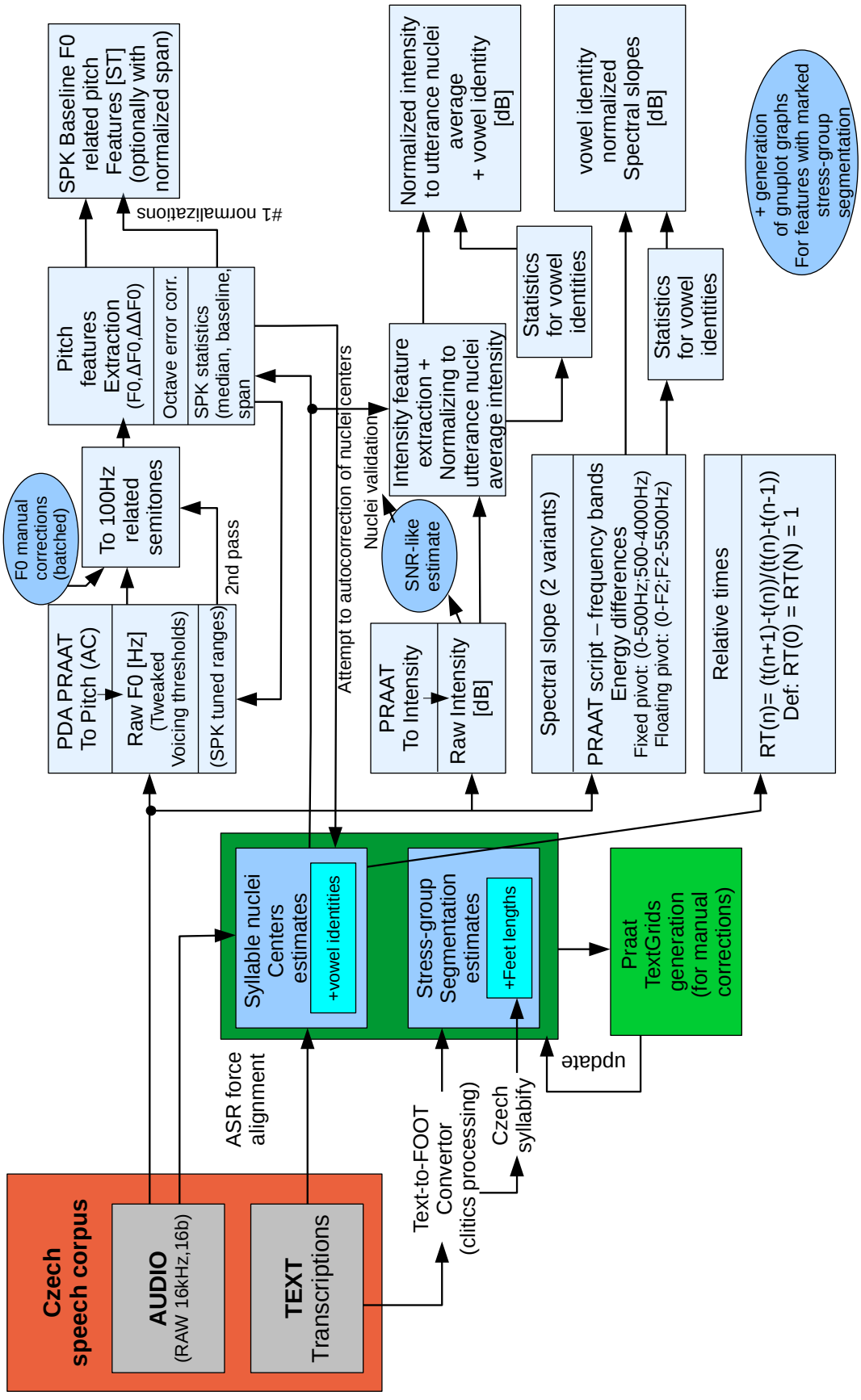


Figure 6.3: Diagram of feature extraction for stress-group segmentation task based on force-aligned syllable nuclei times

are expected to be perceived by the listeners were extracted. According to the previous research, the centers of syllabic nuclei (vowels or syllabic 'R' or 'L' in Czech) were chosen as those pitch representative moments. The syllabic nuclei center timestamps were obtained from force-alignment of reference transcripts using existing Czech ASR system (based on trained context tri-phone GMM-HMM models in HTK environment) as times of middle state occurrence of aligned context tri-phone.

Extracted pitch values (which count corresponded to the count of syllables in each utterance) were then considered as raw feature values and were normalized in 3 different ways (marked as "norm1-norm3", while "norm0" represents the raw features). The first type of normalization ("norm1") is based on the knowledge of utterance division into the feet and relates computed pitch to the mean pitch of given foot. Second type of normalization ("norm2") is based on fitting the "norm0" values with the 2nd order polynomial function and computing the difference of each pitch point to the corresponding function value. Unfortunately, the very last foot, as being often very decreasing in pitch (all the utterances in the data set were of declarative modality), tends to make the curve more convex for the other parts of the utterance. This is why "norm3" was introduced where the same process as for "norm2" is applied but with removed last foot of the utterance from all data structures. Results of described pitch normalizatins can be seen in the figure 6.4.

6.4.3 Model training

To model our problem a Hidden Markov Model (HMM) approach was utilized, because the task is similar to standard speech recognition tasks, where HMM framework is widely adopted. All the experiments were performed using HTK Speech Recognition Toolkit [132]. HMMs for various feet lengths (1-8 syllables) were trained on the training subset consisting of 90% of all the data. The models of stress-groups are in comparison with standard speech tri-phone models without state self-loops and backward state transitions. Thus, the model is not allow to stay in any state and state-flow of our system is strictly forward with coming input features. HMM model for two-syllable stressed-group (label FOOT2) is illustrated in the Fig.6.6a. Each emitting HMM state was modeled using one mean and variance (no mixtures). Each utterance is generally modeled by the grammar depicted in the Fig.6.6b. The a priori probability was not modified for any of the models and thus is equal for all of them. During training and testing, utterances were not chained together to created one long sequence of stress-groups, but rather remained the basic unit for model training/testing.

For all pitch normalization types separate HMM models of Czech feet with different

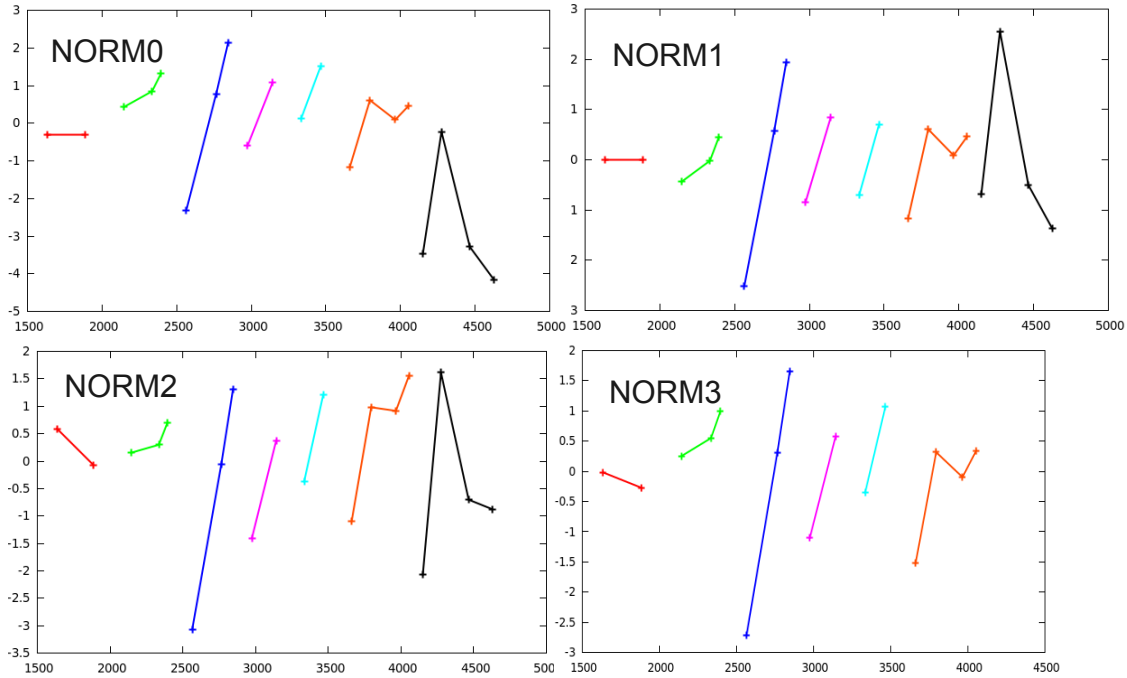


Figure 6.4: Pitch normalization types applied on sample utterance used in GMM-HMM experiment ("norm0" denotes raw input features). X-axis represents time in milliseconds, Y-axis represents pitch in semitones.

syllable lengths were trained. As for testing with "norm1" data, it is clear, that in real system those data will not be available, because the division of the utterance into stress-groups will be unknown. This is why a decision about making experiments with feet models trained on "norm1" data, but tested with different normalization types available in real situation, was done. Also, version of experiment with filtered final feet of utterances was prepared. The expectation was that the filtering could improve the accuracy of the system, because the feet most influenced by sentence intonation would be remove. Besides, in another experiment version all the sentences that contained comma or indirect speech were filtered out, because their feet might be mostly affected by complex sentence intonation.

6.4.4 Results

Various versions of experiment were performed, but only those most valuable are quoted in Tab.6.2. The key finding here was, that neither filtering of the last foot nor filtering the complex-sentences out of the dataset did not improve the accuracy.

Once the optimal alignment has been found, the number of substitution errors (S), deletion errors (D) and insertion errors (I) can be calculated. The percentage correct

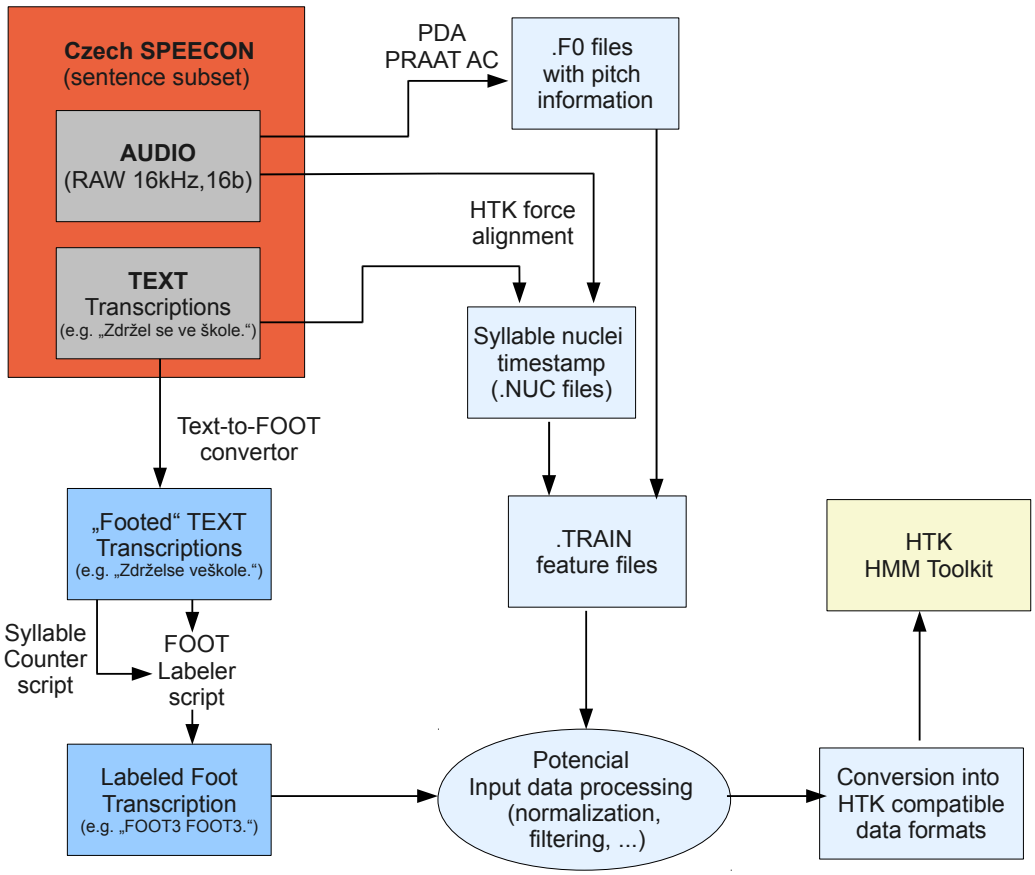
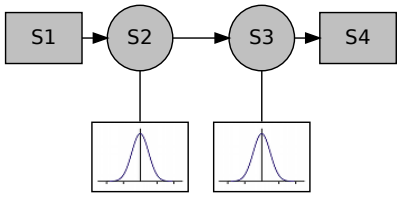
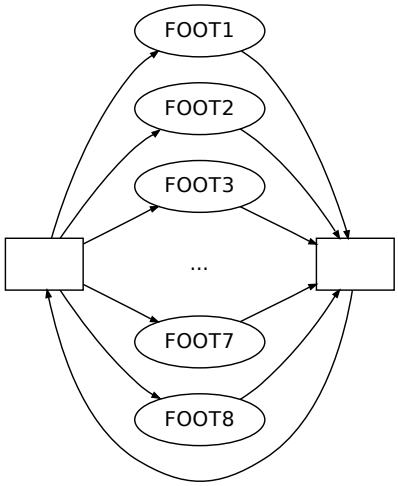


Figure 6.5: Scheme of the training data preparation process



(a) HMM prototype of two-syllable FOOT2



(b) Used Grammar

Figure 6.6: Illustration of used HMM modeling

($Corr$) is then

$$Corr = \frac{N - D - S}{N} \times 100\ [\%] \tag{6.2}$$

where N is the total number of labels in the reference transcriptions. Notice that this measure ignores insertion errors.

For many purposes, the percentage accuracy (Acc) defined as

$$Acc = \frac{N - D - S - I}{N} \times 100\ [\%] \tag{6.3}$$

is a more representative figure of recognizer performance [133].

By using "norm1" type of features, the obtained accuracy was the highest, up to 46%. This denotes a suitability of this normalization type for our task compared to raw feature data ("norm0") with accuracy of 34%.

In the more real scenario using norm3 features as testing data while evaluating the HMM models trained on norm1 features, the system was able to recognize feet in utterances with only 32% accuracy.

In the Fig.6.7 are depicted the "norm1" trained models FOOT1-FOOT7 with plotted pitch mean and standard deviation values of emitting states. After a visual review of trained models (statistically averaged training examples), it can be seen that models for FOOT2-FOOT6 very well satisfy the theoretical condition for foot existence declared in [75] – that intonational contour with pitch drop in the middle does not create acceptable form of foot in Czech. Also, one could see foot shapes with accordance with previous research on Czech stress: initial gradual rise on first few syllables with optional subtle final fall for feet longer than 4 syllables (clearly visible for FOOT5). Even, a clear raise of about 1 semitone between the first and the second syllable can be observed for stress-groups with lengths from 2 to 6 syllables (FOOT2-FOOT6), although the first syllable show also consistently increased standard deviations compared to the rest of syllables within given foot.

Training data	**Testing data**	**Corr**	**Acc**	**D**	**S**	**I**
norm0	norm0	34.2	18.2	2326	2718	1223
norm1	norm1	53.4	46.1	1739	1755	561
norm2	norm2	30.4	18.7	2659	2563	876
norm3	norm3	32.1	20.4	2240	2154	759
norm1	norm3	38.7	32.1	1460	3145	493

Table 6.2: Results from HResults for foot detection using pitch information

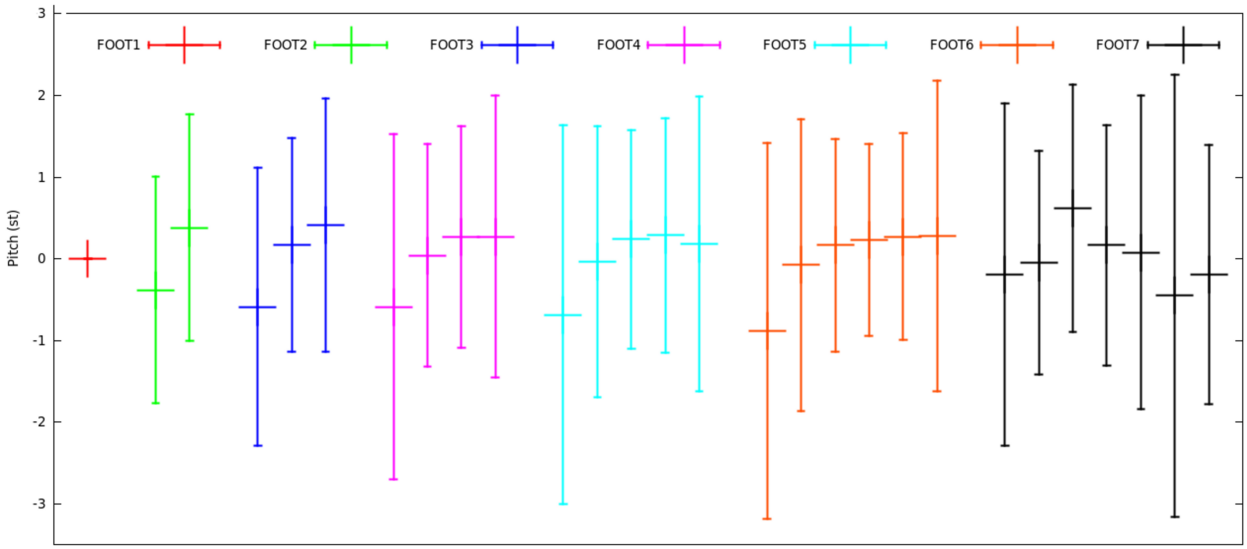


Figure 6.7: Pitch means and STD for “norm1“ trained HMM models of Czech feet

6.4.5 Discussion

Although reached results are not much impressive, I believe that trained model statistically correspond with the most used pitch patterns across various length stress-groups in Czech (and across almost 500 speakers). The low-performance of the system can be explained in various ways:

- There are various pitch patterns varying even for stress-group of the same syllable length, while the model tried to generalize them into the one average realization.
- Pitch itself is not enough as self-explanatory feature for stress group segmentation and the contours for delimitation of the stress-groups must be taken as multi-dimensional, including other acoustic features (intensity, duration and spectral slope, even that it is known that non of them plays a primary role for Czech stress marking).
- The automatic delimitation and annotation of stressed-groups from unprocessed utterance transcripts contains severe amount of errors which might ruin the experiment. Obtaining a verified set of stress-group labels might improve the system accuracy.
- The input phone force-alignment from which syllable nuclei center timestamps were derived is not precise enough to find the nuclei center precisely enough in all the cases and brings an error for pith feature extraction in terms of correspondence with perceived pitch.

Unfortunately, presented modeling technique does not cover the situations which occur at the stress-group boundaries. As I expect, exactly those locations (between the last syllable of previous foot and the first syllable of the following foot) provide a valuable information in terms of acoustic contrast needed to stress-group demarcation. This is why another experiment which is rather syllable context based was carried out further in my research and is covered in the section 6.6.

6.4.6 Precision of automated nuclei centers estimates

Referring to the last point of previous low-performance issue discussion, the precision of syllabic nuclei center estimates obtained from automatic force-alignment by HMM-GMM based Czech ASR system was manually examined. Several randomly selected utterances from used dataset were studied using Praat [112] environment and the precision was explored individually for each vowel identity (distinguishing also between short and long vowels, the long variant is marked with doubled syllable label like 'aa' for Czech [á]). The method of manual syllable nuclei center time determination did not take into account an intensity contour and related only on vowel segment boundaries estimates obtained from subjective listening. Table 6.3 shows those individual statistics together with globally averaged measures.

vowel	a	e	i	o	u	aa	ee	ii	oo	TOTAL
count	26	32	17	17	10	9	2	16	3	136
diff. sum	109	188	92	149	73	106	40	143	22	951
avg. diff.	4.19	5.88	5.41	8.76	7.30	11.78	20.00	8.94	7.33	6.99
STD	8.95	12.66	10.36	13.59	11.35	9.16	0	21.40	6.43	11.78

Table 6.3: Precision of automated syllable nuclei centers estimates for Czech vowels using HMM-GMM based ASR force alignment on the phone level. Positive difference values indicate earlier automatic alignment nucleus center estimates compared to manual method. All values except the counts are in milliseconds.

Czech long vowel [ů/ú] was not contained within randomly selected utterances and thus is missing in the summary tale. Positive values indicate earlier nucleus center estimate from automatic alignment against manual method. Global tendency of the used force-alignment is to consistently align the nuclei centers earlier for all vowel identities and that long vowels have higher estimate differences than the short vowels. Except this global tendency, there are clearly worse estimates for short vowels 'o' and 'u' (high average difference accompanied by high standard deviation values) suggesting that used ASR system might have troubles aligning those vowels correctly. According to the table, the worst estimates of syllable nuclei centers can be expected for long vowel [í] labeled as 'ii'

with average earlier estimate by 20 milliseconds.

Except presented rough quantitative statistics, an objective criterion was also brought in. The time estimate differences were investigated in terms of pitch differences according to the pitch contour obtained by Praat autocorrelation method with default settings. two categories were established regarding the effect on extracted pitch feature values: the first one included serious cases when time nuclei center estimate difference produced the difference in pitch about 1 semitone (or more). The second category included less serious cases covering a pitch difference (typically between 25 and 50 cents, which correspond to values from a quarter to half of the semitone). There were 5 cases out of 132 examined syllables that belong to the first category with particular positive time differences of 50 ('ii'), 50 ('ii'), 40 ('o'), 30 ('e') and 30 ('u') milliseconds, all force-alignment estimates being "early". In terms of less serious cases, only three of them were observed out of 132 examined syllables.

To summarize, although there exist strong tendency of used ASR force-alignment method to place the syllabic center earlier than in manual extraction, there were only several cases when this might seriously influence extracted pitch values.

During this manual verification, a few of halving octave errors were observed with real cause by some degree or kind of creaky voice present in the speech signal, which would be very probably not perceived in non-critical listening like usual conversation.

6.5 Czech stress/unstressed syllables experiments

6.5.1 Intensity of stress/unstressed vowels

Figure 6.8 illustrates computed statistics (mean and standard deviations) of stressed, global and unstressed Czech vowels with intensity extracted in the syllable nuclei centers. 0dB corresponds to average intensity of syllable nuclei centers per each tested utterance. We can see relatively small steps, while there is no clear pattern considering vowel quantity (vowel length; short or long).

6.5.2 Fixed context groups for vowel intensities

The main idea behind presented procedure is a suggestion, that the acoustic intensity of particular vowel in the utterance is very likely to some extent influenced by the neighboring vowels. Thus, different sequences of vowel will very likely behave differently in terms of their intensity values, too. To minimize mentioned phenomena, it was decided to examine the behavior of stressed/unstressed vowels in comparable "fixed" 1-syllable left-

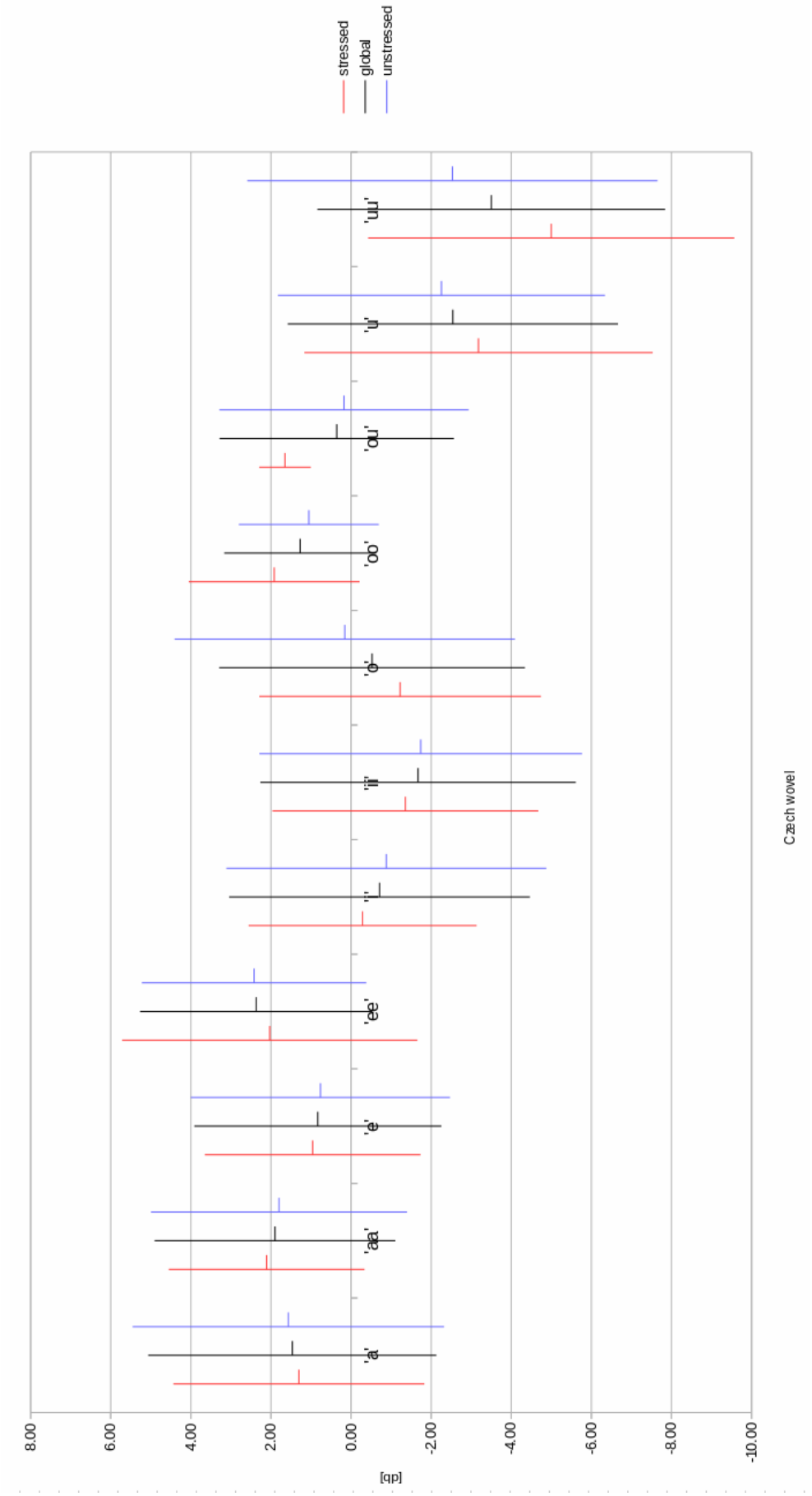


Figure 6.8: Czech vowel mean intensities with stressed and unstressed realizations (results on manually verified clitics-free SPEECON subset consisting of 189 utterances)

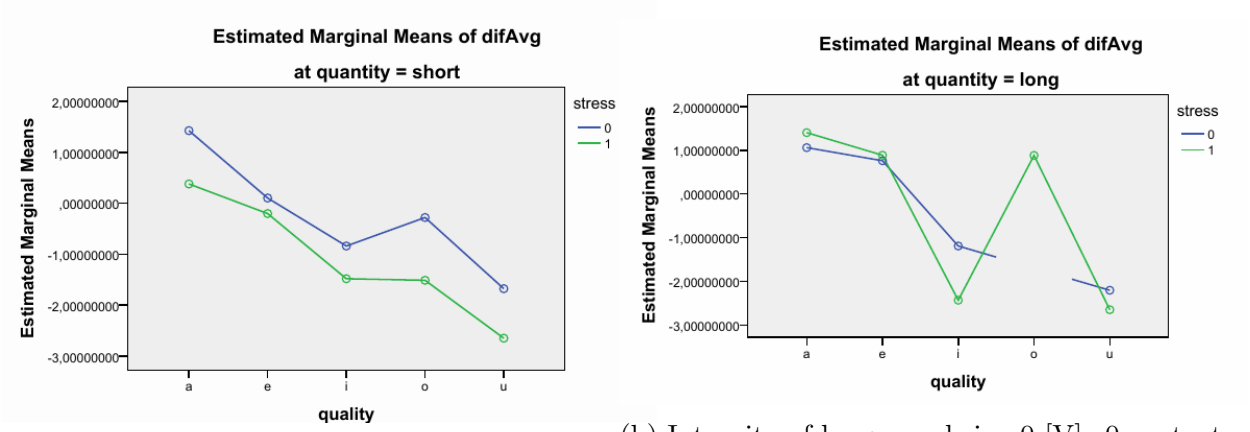
right context. The most common combination found in the SPEECON full subset was "e0-[V]-e0" (short unstressed 'e' followed by vowel V followed again by a short unstressed 'e'). For this fixed triplet context there were 3219 occurrences of our interest across whole corpus. Average intensities of short and long vowels in fixed "e0-[V]-e0" contexts can be seen in the figure 6.9. For short vowel, there is a clear trend for stressed vowels to be less loud about 0.5-1.0 dB than the unstressed vowels (while [e] exhibits the smallest difference). For long vowels in given context there is no overall tendency, but this might be biased by the absence of stressed long vowel [ó] in given context "e0-[V]-e0". Counts of occurrences of individual vowels within given context can be seen in the table 6.4.

Occurences	**Vowel [V]**	**Stress**
48	aa	0
123	aa	1
84	a	0
505	a	1
11	ee	0
30	ee	1
166	e	0
684	e	1
67	ii	0
80	ii	1
170	i	0
230	i	1
18	o	0
636	o	1
19	oo	1
9	uu	0
82	uu	1
49	u	0
207	u	1

Table 6.4: Counts of vowel occurrences within fixed context "e0-[V]-e0"

6.6 Advanced modeling of Czech stress-groups

As following of previous experiment 6.4 it was decided to change the strategy and model individual syllable stress category using context information and more acoustic features. The next difference was to use manually verified data only. Although their magnitude would not reach the set used in previous experiment, it was believed to be sufficient for simple classifiers that were to be tested.



(a) Intensity of short vowels in e0-[V]-e0 context (b) Intensity of long vowels in e0-[V]-e0 context

Figure 6.9: Intensity of short and long vowels in fixed e0-[V]-e0 context

6.6.1 Used features

The set of acoustic features is described briefly already in section 6.3 and comprised of following list extracted at syllable nuclei centers :

- relTimes: relative nuclei times ratios
- pitch: Praat F0, speaker F0 baseline related, fixed octave jumps, pitch values
- pitchVel: Praat F0, speaker F0 baseline related, fixed octave jumps, pitch delta (velocity) values
- pitchAcc: Praat F0, speaker F0 baseline related, fixed octave jumps, pitch double delta (acceleration) values
- intensity: with intrinsically normalized vowels
- SpectSl1: spectral slope with fixed pivot at 500 Hz
- SpectSl1Bp: spectral slope with fixed pivot at F2 Hz

From the recent knowledge about Czech stress group it is not sufficient to try to just distinguish between stressed/unstressed syllables Czech stress has more complex behavior and is more "contour-based" based than (probably) in any other fixed-stress language this leads in the need of context information to be incorporated in its modeling. Particularly according to [121], the context of 2 left/right syllables (so 5 syllables in total) was used as a feature vector with the currently labeled syllable in the middle. This leads to 35 features for one target label, which was stressed (0) or unstressed (1). Used context also also implicates that the first "actual" syllable in the utterance, for which is the context available is the third syllable (first two are not used as target labels) and also the very last

two syllables act only as the context information. Thus, each utterance consisting of N syllables can add only N-4 training/testing instances into the process. The used context will be later denoted as "ctx2 L/R" or "ctx2" (chapter 8).

6.7 Classifier

The task is a supervised learning. CART-style (classification and regression trees) based classifier was found to perform the best from the set of tested classifiers (covering mainly MLP, conditional random fields and SVM).

Dataset comprised of all available manually verified corpora described in section 5.2. As dataset was heavily (but naturally) unbalanced with three times less stressed syllables than unstressed, weighting factor for stressed syllables was increased by factor 3 for the training. This heavily improved the results.

Training/testing set was for all tested corpora divided automatically by 10-fold cross validation. Reported results are averages over all combinations, while model exported for evaluation of other testing set was the one that performed the best.

6.8 Evaluation of the stress-syllable models and results

In the results, TP stands for True Positive, FP state for False Positive. Precision involves TP and FP, while Recall covers TP and False Negative (FN, not present in the binary classification task). F-measure is a harmonic mean of Precision and Recall and is usually the best single overall measure for the model performance.

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.642	0.369	0.799	0.642	0.712	0
0.631	0.358	0.435	0.631	0.515	1
0.638	0.366	0.688	0.638	**0.652**	Weighted Avg.

Table 6.5: Eliška Churáňová corpus (6 speakers), 10 fold-cross validation

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.723	0.278	0.831	0.723	0.773	0
0.722	0.277	0.58	0.722	0.644	1
0.723	0.278	0.744	0.723	**0.728**	Weighted Avg.

Table 6.6: CART SPEECON clitics-free manually verified subset, 10 fold-cross validation

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.735	0.264	0.864	0.735	0.794	0
0.736	0.265	0.548	0.736	0.628	1
0.735	0.264	0.768	0.735	**0.744**	Weighted Avg.

Table 6.7: CART model trained on SPEECON clitics-free manually verified subset, testing set is filtered corpus from diploma thesis of Helena Spilková

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.627	0.405	0.782	0.627	0.696	0
0.595	0.373	0.408	0.595	0.484	1
0.617	0.395	0.669	0.617	**0.632**	Weighted Avg.

Table 6.8: CART model trained on SPEECON clitics-free manually verified subset, testing set are 4 speakers from Eliška Churáňová corpus

6.9 Discussion

Lower performance of Eliška Churáňová corpus can be explained by current inability of used modeling technique to capture correctly the stress-groups containing anacrusis (unstressed syllable) as the first syllable in the foot. Unfortunately those are quite common in the broadcast news speech and were not removed, because of their relative occurrence. The removal would lead to not using whole utterance, which would decrease the corpus significantly. On the other hand, the SPEECON clitics-free manually verified subset seems to match the filtered corpus from diploma thesis of Helena Spilková, which is marked by the achieved F-measure of 74.4%.

Chapter 7

Sentence/phrase modality classification

Sentence modality classification task very closely relates to punctuation insertion task or sentence boundary detection task. One approach for solving punctuation insertion is that ASR system tries to find a suitable slots for punctuation marks insertion, which can be retrieved by sentence or phrase boundary detection. The segmented sentences or phrases can be further investigated for their modality by the modality classification module and the finally, punctuation mark corresponding to given classified modality is then inserted in the place of the slot. All of mentioned stages might use lexical or acoustic information or their combination.

7.1 Usage of modality detector

A modality detector can serve as punctuation insertion module. Text structuralization with punctuation marks on the output of ASR system is crucial for its readability and also for the case of following machine translation into another language. Although, there are clear cases which illustrate the lack of correspondence between punctuation and prosody in both directions [134], it can be claimed that there is still not negligible amount of cases, when there is clear correspondence between them. At least for phrase melody, its modality and dedicated punctuation mark that express this modality in written form seems to be this kind of correspondence.

In dialogue systems, there is need for retrieval the semantic of the user's inputs. Modality detector serves in this kind of system as an adviser in the natural language understanding process by delivering information about phrase modality (whether it was a question or statement in the user input). In this task, it is not insisted on perfect

punctuation marks mapping from the phrase, but it is dealt with pure phrase modality classification, with emphasize on distinguishing between questions and declarative sentences.

7.2 Related works

Currently, automatic punctuation is very often limited to commas and full stops insertion.

There exist approaches using pure lexical information to accomplish the task. In [36] conditional random fields were used to do better punctuation predictions, while missing punctuation restoration using LSTM RNN was presented in [135].

Recently, punctuation prediction task was also the subject of "monolingual" machine translation (MT) [136], where ASR output without any punctuation enters the MT system which translates it into the same language but this time enhanced with punctuation marks (compared to classical MT from one language into another). Used approach brings additional latency to the real-time system due to additional context needs compared to standard methods, but even those cons were partially addressed [137].

A solution in current Czech transcribing systems how to implementing dictation task enhanced with punctuation marks in real-time (or near-real-time) is either their manual insertion or dictation of predefined punctuation key-words (personal experience with current commercial Czech dictation systems). Particularly, live captioning (subtitling) task for Czech is being described in [138] using professional re-speaking of broadcast content into ASR system. A re-speaker uses his/her hands to press punctuation marks on a keyboard during inter-word pauses, so they can be processed by the recognition system that presents the punctuation marks directly in its result.

The first experiment on assignment of sentence type with respect to the speaker's standpoint for Czech was carried in [139], where artificial neural network was used for the classification. Each sentence was divided into N windows and in each of the window voice energy and the fundamental frequency were extracted as features. This means there was time "compression" for longer sentences as the input to the network was fixed to 2N. Reported results show 83% correct classification out of 350 testing sentences of four different modality. No information is although provided about training and testing set relativity. It might be assumed, that the training set was the same as the testing one, which would explain that high accuracy of the system.

In [140] Klečková suggests three-tier punctuation hierarchy according to the pause duration, for modality related punctuation marks the pause duration occupies interval of 200-240 ms.

In [141] the sentence boundary detection was inspired by [142], but the final punctuation mark classification was added. The task of the paper was the punctuation annotation from read speech using prosodic and lexical information, while examining those trained models separated and in combination. The principle was, that for each inter-word boundary one of the following targets was assigned: comma, sentence boundary and no punctuation. The target sentence boundary joined full-stops, question marks into one category because of lack of the latter in used corpus.

A prosody model consisted of word-level based contextual features: two preceding words, actual word and one following (4 words in total). Author mentions an individual prosodic speaker style and mood and need of normalization and smoothing techniques for obtaining valid features. Direct modeling strategy was adopted for feature extraction from the automatically aligned speech signal. Word-level based prosodic features were related to F0, phoneme duration, pause lengths and energy. Two prosodic classifiers were trained: CART decision tree and MLP neural network. F-measure for both prosodic classifiers/detectors was in the range of 62-63% and raised to 65-66% using modified F-measure score (giving half score when punctuation place is correct, but punctuation symbol is wrong).

One of the parts in work [143] also deals with Czech punctuation insertion. It restricts the insertion to commas and full-stops, while it combines information from lexical, acoustic and morphological level to mark the slot for punctuation placement. The acoustic information is here limited to "noises" detected from ASR (typically silences and breathins) and thus any true prosodic features are not involved in the process.

In work [144] more acoustic features are utilized compared to any previous works, but still the prosodic unit for which single features are derived is a word. They focused on estimation of commas and full stops only in Czech and Slovak radio broadcast archive. Commas estimates were based on ASR text output only. Special N-gram language model was used in combination of derived rules for allowed (positive rules) and forbidden phrases (negative rules) both before and after the comma punctuation mark. There were more negative rules compared to the positive ones fro Czech.

Full stops estimates were on the other hand based on textual, prosody and document segmentation cues. Used prosodic features involve F0 (5 candidates were extracted by STFT and chosen by dynamic programming decoding/smoothing) and non-speech events.

Authors describe that the last word in the sentence has usually very changing pitch while the words inside the sentence keep flatter pitch trend. This is why two F0 related features were extracted: the mean F0 of the word and normalized difference between maximum and minimum F0 of the words (characterizing the F0 variation to some extent). The sentence borders were detected using the scheme, where mean pitch of subsequent word pairs declines and the normalized pitch difference of the second word exceeds a given threshold. If in the proposed sentence border is not already suggested a comma, the ASR lexical output around the slot is searched for forbidden words (that cannot stand in the beginning or end of the sentence) and if this is not the situation, the full stop is placed. Results for prosody assisted full-stop detection shows high 80% precision (if mark was placed, it was placed correctly in 80%), but very low 25%recall (only 25% of slots were marked) by the detector.

Compared to the related works, a classification modality approach presented in this thesis relies contrarily purely on acoustic information. From the knowledge of theory about Czech modality system, where the last stress-group is bearing melodeme (which determines the modality) in the most of Czech neutral phrases/sentences, the task might be in specific use-case and point of view quite similar to the stress-group segmentation. Again, information about the syllable nuclei centers can be advantageously utilized for achieving the most close approximation to human perception of nuclear pitch accent when the pitch information is extracted in those "peaks" of syllable sonority.

7.3 Experiment variants

In practice (real use-case) we might face several options for available acoustic information of phrases:

- What features are known?
    1. "completely uninformed" - unknown last foot (melodeme) scope, unknown syllable nuclei center timestamps. This kind of experiment was carried out initially and is described in [145]
    2. unknown last foot (nuclear pitch accent) scope, known syllable nuclei center timestamps series
    3. known last foot (melodeme) scope, unknown syllable nuclei center timestamps series
    4. "completely informed" - known last foot (melodeme) scope, known syllable nuclei center timestamps series

- Are speaker pitch statistics known?

    1. unknown speaker's pitch range, unknown baseline F0: we can normalize only to utterance average
    2. unknown speaker's pitch range, known baseline F0: we can normalize to baseline F0
    3. known speaker's pitch range and baseline F0: we can normalize according to both statistics. Very beneficial behavior is expected because of the ability to fine position speakers pitch sequence into his overall pitch speech scope.

- Separate isolated phrases or flowing stream of speech?

    1. classification of separate prosodic phrases (each corresponds exactly to one nuclear pitch accent)
    2. detection and classification of punctuation marks from data stream:
       Training/testing data consist of stream of discrete pitch values in nuclei, each value is followed by the target punctuation mark. Target set is in this case might be one of [noPunct , . ? !]. Becoming a detection problem, used evaluation metrics are then True/False Positive/Negative and derived precision, recall and F-measure.

7.4 MLP experiment on Czech modality

There can be found mapping between pitch patterns and modality classes in Czech. From the theory described in section 3.4.3, this mapping is not unique as concluding descending type stands for at least two punctuation marks, and there are two nuclear tones for the question mark. Those facts make the task more difficult.

The approach used in presented experiment [145] can be viewed as a basic feasibility study of prosodic "standalone" automatic punctuation detector for Czech language. "Standalone" property means that the module can be almost independent on hosting ASR system, because the punctuation detector will not use any of the information provided by ASR (recognized words and its boundaries, aligned phonemes duration, etc.) and will operate directly on raw acoustic data.

To illustrate the problem, an example of raw pitch pattern of Czech complex sentence is depicted in the figure 7.1.

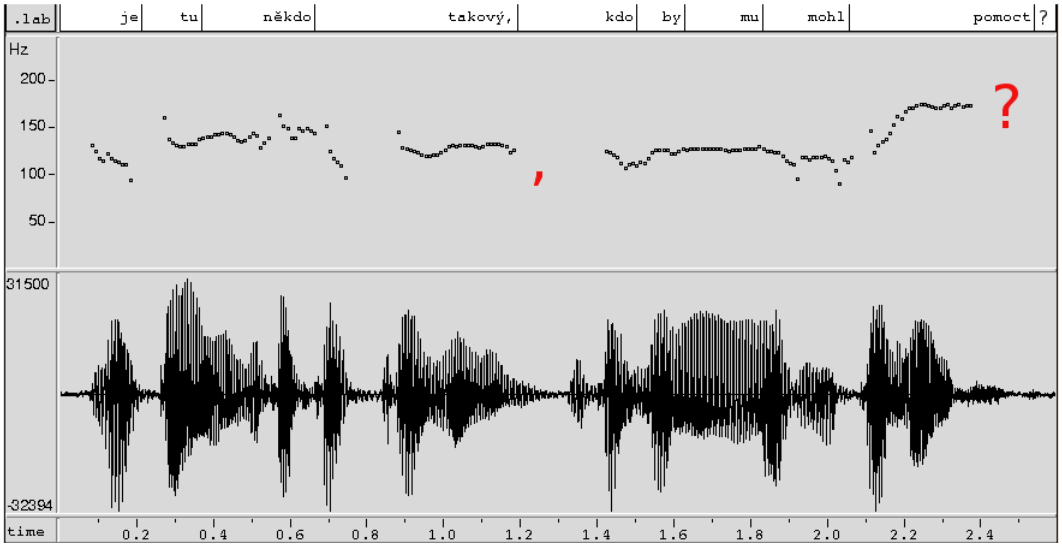


Figure 7.1: Pitch pattern of Czech complex sentence "Je tu někdo takový, kdo by mu mohl pomoct?" ("Is there anybody, who could help him?"). The second part of the sentence is dependent on the first part.

7.4.1 Time series pattern detection

The problem of detecting patterns in time series has been widely investigated and deals mainly with the amplitude variability, time scale variability and absolute value shift of the observed patterns.

There are studies that try to appropriately pre-process the time series in the scope of a sliding window and than run matching algorithm to compute distance from the searched patterns in defined metrics [146], [147]. On the other hand, an artificial neural network (ANN) approach for finding patterns in time series was also developed in the past, especially by Elman [148] network type. [149] brings nice overview and introduction into problems on either conversion the time domain into spatial one or utilization of memory (loopbacks) in network architectures. An example of the application of ANN approach could be [150] utilizing classical multi-layer perceptron and FIR based network or [151] dealing with financial stock time-series data.

7.4.2 Neural Networks for Temporal Processing

Artificial neural networks (ANN) are a well known tool for classification of static patterns, but could also be a good model for dealing with the time series. From theory ANN could be seen as non-linear statistical models. Multi-layer perceptron (MLP) networks can be considered as a non-linear auto-regressive (AR) model and can approximate arbitrary function with arbitrary precision depending only on the number of units in the hidden layer. By training the network we are trying to find the optimal AR-model parameters.

Two basic approaches for the classification of temporal patterns are: 1) a usage of the classical MLP feedforward network or 2) usage of special type of neural network with 'memory'. In the first case we are dealing with a fixed number of inputs in the input layer of the network, where no 'memory' is available. This means we need to map time dimension onto spatial one by putting the whole fixed-length frame of signal onto all the inputs of MLP network. The main issue is that time patterns vary not only in amplitude, but also in its duration and thus we need to choose suitable frame length. In the next step another frame (depending on the shift of frames) of signal is brought on the inputs and the network gives a new answer with no connection to the previous one. In the recurrent types of network, there are loopbacks creating the memory. This architecture allows us to have only one input and bring one sample on it in each step and get new output.

7.4.3 Training and Testing Data

Although there are many databases for the training of ASRs, not many of them can be used for our task. Firstly, in most of the cases punctuation marks are missing in the transcriptions in these databases. This flaw can be removed by re-annotating the data and putting punctuation marks back in the right places. Secondly, there is often a shortage of prosody and modality rich material in these databases. And what is worse, if the material exists, the speakers in most cases do not perform the prosody naturally, because of the stress when being recorded. Special emotive databases exist too, where certain parts of it can be used, but emotions of speaker are not the object of our study. That is why alternative data sources were looked for.

Finally, the online library of Czech audiobooks read by leading Czech actors was used. A compressed MP3 format of audio files did not seem to be an obstacle as the records are very clean with studio ambient. In addition, actor's speech is a guarantee of intonation rich material. For first experiments presented in this paper a basic sample of the library including unified data from 4 different audiobooks read by 4 different actors (3 men, 1 woman) was manually annotated to roughly include a natural ratio of punctuation marks for Czech language. The counts of individual punctuation marks can be found in the table 7.1. As in the future we plan to increase the amount of data with use of an automatic alignment system based on available electronic versions of the books, we did not manually mark the places where beginning of the cadence occurs. This task would even need phonetic specialists assistance and it is very difficult to automate. That is why intonation pattern for corresponding following punctuation mark is taken from the beginning of the whole sentence or previous non-concluding punctuation mark (comma). Basic intonation contour was computed directly using PRAAT [112] cross-correlation PDA

with default settings.

Punctuation mark	**?**	**!**	**,**	**.**	sum
count	31	7	65	158	261

Table 7.1: Occurrences of punctuation marks in the used data set

7.4.4 Pre-processing the Intonation Patterns

Raw data pre-processing is a common first step to meet requirements of the task. When using the neural networks for pattern classification, there is also need to prepare the data to maximally fit the chosen network architecture.

1. **Logarithmic scale conversion**
   Due to the fact that a human perception of pitch occurs in roughly logarithmic scale, we need to convert frequency values (in Hz) into musical scale values (semitone cents) according to equation 7.1, where ideally f_{LOW} is a low frequency border of vocal range of the speaker. This makes the signal values relative to this threshold and deletes differences of absolute voice heights (curves of same patterns should now look the same even for man or women speech). This conversion also implicitly removes the DC component of intonation signal, but it also means we need to know what the lowest frequency border of the vocal range of the speaker is. From training and testing data sets this can be computed as finding minimums over all of the units spoken by the speaker. When applied online, we will gradually make the estimate of this value more and more accurate.

$$Cents = 1200 \log_2(\frac{f}{f_{LOW}}) \tag{7.1}$$

2. **Trimming the edges**
   As the annotated patterns have silent passages on the beginning and at the end (zero-valued non-voiced frames), we need to remove these parts of the signal for further processing (see the next step).

3. **Interpolating missing values**
   The speech signal consists not only of voiced frames when the glottis do pulse with certain period, but also of unvoiced frames when the glottis do not move (unvoiced consonants). Good pitch detection algorithm can distinguish between these two cases. This leads to situation of having zero values as a part of the intonation curve. Such a curve does not seem to be continuous even for very fine time resolution.

Because these "zero moments" depend on certain word order and not on supra-segmental level of sentence, we need to get rid of them and thus maintain that same intonation pattern with another words in it leads to the same final continuous intonation pattern.

4. **Removing segmental differences of intonation**
   As we are following intonation as supra-segmental feature of speech, we are not interested in intonation changes that occur on intra-syllable level. That is why we want to erase these finer nuances and maintain only the main character of the curve. This can be accomplished by applying an averaging filter on the signal. We could also achieve similar result by choosing longer signal window and its shift in pitch detection algorithm setting.

5. **Reconstructing the levels of extremes**
   Previous smoothing unfortunately also smoothed out the intonation extremes, changing their original pitch. Because these extremes are very important for pattern character, we want to 'repair' them. In current implementation only two global extremes are gained to their former values by adding (subtracting) appropriately transformed Gaussian curves with height of differences between original and smoothed values and with width of previously used smoothing filter.

6. **Signal down-sampling**
   High time resolution of time patterns leads to a need of a high number of inputs for classical MLP or long 'memory' for recurrent ANN. Both facts imply a higher unit count in both types of network, which is then more difficult to train with limited amount of training data. That is why down-sampling of pattern is needed. Down-sampling is done several times according to the type and architecture of ANN used for follow-up classification:

   (a) 'Normalizing' down-sampling - MLP type of network with temporal into spatial domain conversion needs fixed length vector on its input. Each pattern is thus normalized in its length to satisfy the 64 or 32 input vector length condition.

   (b) Classical down-sampling - recurrent networks do not require fixed-length patterns, but to perform reasonably, too precise time resolution of the series implies high count of hidden units . That is why a classical down-sampling from 1000 Hz sampling rate to 40 Hz and 25 Hz is done.

7.4.5 Results and Discussion

The experiments were made on the data set, where 70% of it were training data, 15% validation set and 15% test data. Trained network was then evaluated on the whole data set. Results for intonational patterns with fixed length of 32 samples on MLP with 15 units in hidden layer can be seen from the confusion matrix (table 7.2) evaluated over the whole data set. The classifier tends to prefer classes with higher occurrence in training data set (commas) due to their statistically higher occurrence in validation set. That is why an another experiment was done using a limited equal distribution of the patterns in the classes (all the classes contain 31 patterns except the exclamation mark class). Representative confusion matrix for the reduced data set is in the table 7.3 for 32 samples per pattern and 20 hidden units. From the results it is obvious that the MLP network is capable to give near a 50% success classification rate for classes of question marks, commas and full stops. The impossibility of classifying exclamation marks could be based on the fact that these intonation patterns are not stable in intonation and that this type of modality rather lies in another prosodic feature (energy), or the data set for this class was too small in our corpus. The last experiment was done on cut-length patterns, where only last N={1500,1200,800,500} ms were left, then down-sampled to 32 and 64 samples for MLP input. 64-sample patterns were more successfully recognized. Best results for the reduced data set were obtained for 1200 ms patterns and 10 hidden neurons (table 7.4).

Table 7.2: MLP Confusion matrix in %, full data set, full pattern length

Actual class → **Predicted class ↓**	**?**	**!**	**,**	**.**
?	**10**	0	2	1
!	0	**0**	0	0
,	20	29	**18**	6
.	70	71	80	**93**

Table 7.3: MLP Confusion matrix in %, reduced data set, full pattern length

Actual class → **Predicted class ↓**	**?**	**!**	**,**	**.**
?	**40**	32	32	24
!	1	**4**	1	1
,	33	41	**48**	23
.	26	23	19	**52**

Table 7.4: MLP Confusion matrix in %, reduced data set, last 1200 ms of pattern

Actual class → **Predicted class ↓**	**?**	**!**	**,**	**.**
?	**46**	30	26	28
!	1	**7**	1	1
,	28	40	**57**	22
.	25	23	16	**49**

7.4.6 Experiment conclusion

We discussed two approaches for the classification of sentence modality based purely on the intonation. MLP based approach gives classification success rate around 50% on question mark, comma and full stop classes. As expected, questions are often misclassified as statements. A solution to better results on questions might be to separate them into two classes (open and closed questions), but this definitely leads to higher demand on training data magnitude, which is the problem of presented experiment anyway.

Chapter 8

Integration of prosody into ASR system

There are various options to integrate prosodic information into ASR system at various levels. Here, the focus is concentrated mainly on integration of stress-group detector as this is very valuable information related to word boundaries. Also, possibilities of phrase modality classification system integration are discussed in this chapter.

8.1 Integration of stress-group detector

Depending on the information available from the ASR, we have various options for utilization of stress-group segmentation module output: prosodic evaluation of single hypothesis, prosodic evaluation of N-best hypothesis (which is just generalization of the previous) or ASR lattice rescoring at the lowest level. In all cases, the key for any improvement of the ASR results is that ASR system has its decoder setting adjusted (decoder and the search is "open" enough) so the correct hypothesis appears in the N-best output or are contained in the recognition lattice.

8.1.1 Single utterance prosodic evaluation

The main idea behind real usage on unknown testing data is that it is possible to evaluate each utterance in terms of its prosodic "footing" score, i.e. how well the output of stress-group trained classifier matches the hypothesis suggested by ASR. It can thus support the hypothesis or notify the ASR system about 'weird' hypothesis, which can be understood as prosody supported ASR error detection. The illustration of comparison between stressed/unstressed syllable chains obtained from stress-group classifier and ASR hypothesis is depicted in figure 8.1.

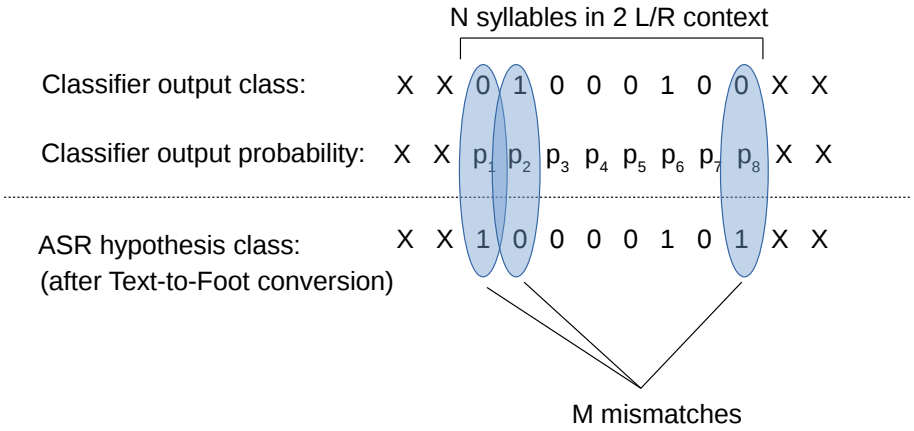


Figure 8.1: Comparison of classifier output with ASR output of stressed (1) and unstressed (0) syllables on $ctx2$ utterance features

Based on our previous research of Czech stress group segmentation, there are basically two kinds of inputs that are needed to compute utterance prosodic score for Czech:

1. audio signal with the information about syllable nuclei center timestamps (the syllable nuclei information being specific requirement for Czech)

2. hypothesized stress-group segmentation of the utterance (syllables marked as stressed or unstressed). As there might be available only ASR hypothesis in common word-based textual form, a way how to obtain stress-group segmentation is utilization of the utterance using Text-to-Foot converter (section 6.2) module again in this place.

3. individual syllable-based stress class output of the prosodic classifier using model trained on acoustic features extracted from audio signal in syllable nuclei centers

Suggested measures of utterance prosodic "footing" score

Two measures are presented for evaluation of single utterance in terms of prosodic segmentation into stress-groups. The first is rough measure $PFSC$ which does take into account only the number of differing syllable labels related to the count of all syllables.

$$PFSC = \frac{N - M}{N},$$

where N is the number of syllables in the utterance and M is the number of mismatches between classifier output and ASR hypothesis.

The second measure $PFSC_w$ presents weighted form of $PFSC$ and value of this score are aimed to be higher (more optimistic estimates). It is applicable only in the case that the output probabilities are provided by the given classifier.

$$PFSC_w = \frac{N - \sum_{i=1}^{M} p_i}{N},$$

where p_i is the output probability of classifier for each mismatch M in addition to definition of $PFSC$. Compared to this definition, the previously defined $PFSC$ can also be understood as special case of $PFSC_w$ with unit mismatch weights. Thus, $PFSC \leq PFSC_w$.

8.1.2 N-best evaluation

Typically, on the ASR output it is common to obtain not only the single overall winning hypothesis, but also the rest of competing hypothesis, in which sometimes the correct one can be found (although it has not been chosen as the winning one, typically thanks to its particular atypical realization or other specific factors). The already sorted list of N-best hypothesis coming from examination of ASR lattice can be also extended by confidence measures per each of hypothesis (or even per each word). Evaluation of all the top N hypothesis coming from the ASR is only a generalization of previous single utterance prosodic scoring. Each hypothesis is prosodically evaluated separately and after sorting their achieved prosodic scores, the prosodic winner(s) can be chosen. Generally, there are various scenarios how to deal with obtained prosodic N-best scores:

- If the ASR did not provide any confidence information of particular N-bests, the two options come into question. If the prosodically winning hypothesis match the very first hypothesis in N-best (which is typically already sorted by final ASR lattice score), it just confirms the top ASR hypothesis. If the prosodic winner differs from the N-best very top hypothesis, informed arbiter should come into play and decide whether prosodic score is high enough to judge the prosodic winner as new winner in overall recognition process or not (then, some additional and more detailed processing might occur which might resolve the situation). Such kind of arbiter awareness comes from the knowledge of reliability threshold of particular trained prosody footing classifier.

- In the case of providing one global confidence measure ASR_{conf} per each of N-best hypothesis by ASR, it can be combined with the computed prosodic score (e.g. $PFSCw$) using empirically obtained constant k (which minimizes final WER) to obtain final utterance hypothesis score SC_{hyp}, like in the equation 8.1. For $k = 1$, the ASR confidence score ASR_{conf} (if scaled as probability in ranges $< 0; 1 >$) is of the same importance as the weighted prosodic score $PFSCw$. More focus on prosodic score can be obtained by setting $k < 1$ and vice-versa, to remain ASR_{conf}

to be the more important part of final score (e.g. for poorly performing prosodic models), setting $k > 1$ can be recommended.

$$SC_{hyp} = kASR_{conf} + PFSCw \tag{8.1}$$

- In the case of providing single confidence measures per each word of each N-best hypothesis by ASR, the task becomes equivalent to lattice rescoring with combined acoustic and lexical scores into single ASR score.

8.1.3 ASR lattice rescoring

The lattice rescoring is the most demanding, but also the most efficient option of integration of prosodic stress-group information into ASR system, because it precedes the procedure of N-best selection. Firstly, standard lattice evaluation is presented, while an algorithm force its prosodic rescoring for Czech is proposed speedily.

Standard flow of ASR lattice scoring

A standard flow of ASR lattice scoring can be described in following steps:

1. On the input we have ASR lattice as the output of ASR decoder graph search. It represents all ASR hypothesis that pass defined thresholds (tweakable) during decoding of the parametrized audio signal of the unknown utterance through the ASR graph (decoding network).
   For example, widely used HTK [152] tool uses the SLF lattice format consisting of:
   - set of nodes: points between hypothesized words with time information
   - set of edges: oriented connections between nodes representing individual words of hypothesis with separate information about acoustic score (word likelihood generated by the acoustic model) and lexical score (word probability within used language model)

   Real lattice visualized from HTK SLF lattice format can be seen in the figure 8.2.

2. Scaling of the lattice:
   Usually, the next step is to scale the AM and LM scores using a given mutual ratio (often defined by a single constant as language model weight or acoustic scale).

   It is also usual to search the optimal value of this ratio so that after final scoring of the lattices they produce overall set of hypothesis (best paths) typically over various utterances with the global lowest Word Error Rate (WER [%]) compared to the reference transcriptions.

Figure 8.2: Real SLF lattice for ambiguous utterance in terms of stress-group segmentation. The real and correct transcript should be "pánové se strojili podle ..." ("The men were dressing according to ..."), while the winning hypothesis contained in this lattice is "pánové sestrojili podle ..." ("The men constructed according to ..."). Acoustic likelihood of link is marked by '-a' option, general language model likelihood of link by '-l' option.

3. Best path(s) searching:
   On the scaled lattice the best path (or N best paths as 'N-best') is (are) finally found representing the 'best' (cheapest) possible sequence of words for given audio signal, acoustic model, language model and their scaling factor. In the last step it is also possible to align obtained best path against audio signal and get time marks if those were missing in the original lattice format (e.g. Kaldi approach).

Modified flow of lattice scoring with prosodic rescoring

The key difference compared to the standard flow of lattice scoring is that next to the acoustic score and language model score we need to evaluate each edge (word) of the lattice graph in terms of its prosodic score. Then, we should be able to combine all three scores (using suitably scaling between all of them) to obtain e.g. best possible WER[%] on the development set of data. The key point here is to be able to evaluate the edge in terms of its prosodic probability. Here we describe one of the possible approaches to realize this (with special attention to Czech clitics system):

1. Assign to every edge a 'neutral' prosodic score. This should be ideally the score, that does not influence the result of best path search (the result should be the same with or even without presence of prosodic score).

2. Prosodic "rewards":
   For each node of lattice, compare its time-stamp with the probability of border at that time from prosodic stress-group model.[1] If this probability exceeds the border acceptance threshold (thr_{acc}), increase the prosodic score of all the edges coming to and from this particular node. There are no other conditions to be met for proper rewarding.

3. Prosodic "punishments":
   Again, for each node of lattice, compare its time-stamp with the probability of border at that time from prosodic stress-group model. If the given border probability is under the rejection threshold (thr_{rej}), the examined word border in lattice can be considered as false placed and automatically punish all the edges coming into and from discussed node. This is particularly true except the two situations, specific

[1] Considering the presented classifier output as depicted e.g. in the figure 8.1, which classifies each syllable as stressed or unstressed, a valid stress-group boundaries can be derived from nuclei centers syllables marked as stressed corresponding roughly to onset times of those syllables as in Czech there is only one syllable bearing the accent in the stress-group and it is (up to the anacrusis) the first one. Probability of border can decrease linearly/exponentially with the time distance from the stressed syllable onset.

for Czech (and other fixed-stress language with plentiful use of clitics). Before we punish the edge we need to check that:

- a) the word on the edge rising from the discussed node (so the word following the word-boundary) cannot be a valid enclitic
- b) the word on the edge pointing to the discussed node (so the word preceding the word-boundary) cannot be a valid proclitic (preposition, ...)

If a) or b) is true, we will not punish this edge as we can consider those situations as valid hypothesis. Otherwise we will lower the prosodic score of the tested edge. Also, for each edge directly containing within itself a moment of proposed stress-group border (the edge connects two nodes whose times present time interval boundaries and the time of proposed stress-group border lies within this interval, possibly with small toleration) we lower its prosodic score.

The particular thresholds and constants need to be tuned with respect of the range of existing acoustic/lexical scores or likelihoods.

To summarize, for Czech, we want all the stress-group borders suggested by prosodic module to be also fulfilled in the lattice (as word-boundaries in a good hypothesis). But on the other hand, we cannot automatically reject the pending hypothesis which has word-boundary in the place that is not being 'marked' by prosodic module by valid stress-group boundary due to the often proclitic and enclitic presence in Czech.

8.2 Possible integration of modality classifier

For successful integration of pure phrase modality classifier into any real system, one prerequisity needs to present, which is phrase boundary detection. For Czech, this field was deeply studied in [54]–[56], while especially section 3.2.2 (Signal-Based Approaches Not Using Textual Information) from [55] is relevant to our lexically blind "approach", and will not be thus covered here in more detail. On the other hand, if modality classifier module can at the same time act as a detector (it follows the sequence of prosodic features and outputs the proposal if suspicious sequence of prosodic features occurs - typically covering the decisive modality pitch pattern), then the phrase boundary detector might be omitted or can serve as additional information for final decision logic. The overall diagram for the use-case of punctuation detector within ASR dictation system can be seen in the figure 8.3. The punctuation detector block works in parallel to standard ASR system. It consists of voice activity detector (VAD), followed by a phrase boundary detector (which

might be implemented according to [55]), that typically searches for suitable pauses in the signal, where a placement of punctuation mark (PM) is valid. It also holds the information about already inserted punctuation marks (and especially the time passed from the last valid PM insertion). The prosodic feature extraction is the input for the main modality classifier block, which can operate in two modes according to the trained machine learning method - either as a classifier or as detector with classification.

In the first variant, it uses the information from phrase boundary detector for end-pointing the input feature vector sequence, which is followingly classified. The punctuation mark corresponding to the most probable phrase modality is passed to the ASR system.

In the second case with ability of detection, it continuously listens to input feature stream and can emit punctuation mark suggestion in any moment. The decision logic (in coordination with phrase boundary detector) than needs to resolve if it is the valid suggestion or rather false alarm from the punctuation detector. Again, the punctuation mark is passed to ASR system for valid punctuation output, where it is interleaved into ASR hypothesis.

The ASR system needs to be modified only in a way, that it can handle signalization from punctuation detector for inserting the suggested punctuation mark into current hypothesis.

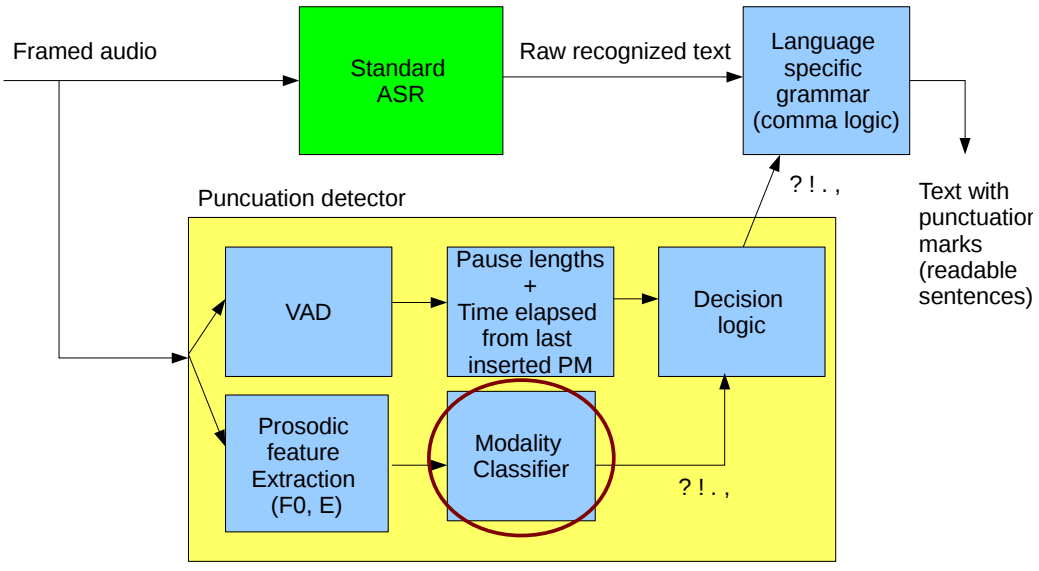


Figure 8.3: Overall punctuation classifier/detector block diagram

8.3 Preliminary results on stress-group segmentation

An experiment based on evaluation of ambiguous utterances in terms of stress-group segmentation by suggested prosodic footing scores was carried out. The testing set was

Label set	Utterance	Hypothesis	Stress-groups	Stress labels
RealLabels	lk001S72	pánové sestrojili podle	3 4 2	1 0 0 1 0 0 0 1 0
AlterLabels	lk001S72	pánové se strojili podle	4 3 2	1 0 0 0 1 0 0 1 0

Table 8.1: Example of testing ambiguous utterance ID "lk001S72" with its real (and such correct) realization of stress-groups 'RealLabels' and with its virtual alternative stress-group segmentation (not realized) 'AlterLabels'. Note, that individual word "se" in the alternative hypothesis transcript is an enclitic joining the previous word "pánové" and creating 4-syllable stress group.

derived from Helena Spilková diploma thesis corpus [76] with 130 filtered utterances (4 speakers, 38 unique transcripts, utterances with non-matching perceived realization with the segmentation aim of the speaker were removed from the set). The complete list of tested utterances can be found in appendix C. The experiment corresponds to ASR 2-best prosodic evaluation.

8.3.1 Distinguishing between real and alternative hypothesis

This experiment consisted of comparing prosodic scores defined in section 8.1.1 for two variants of utterance stress labels. The utterance set in dataset [76] is ambiguous in terms of possible segmentation into stress-groups and there are always two options how the utterance can be segmented. Therefore, for each utterance the real syllable stress labels (that were realized in the given utterance audiofile, mark this set of labels as 'RealLabels') versus its imaginary non-realized alternative (marked further as 'AlterLabels') were proposed. The example of sample utterance is in the table 8.1.

Due to the fact that ambiguous chain of syllables did not cover the whole utterance, but the ambiguity was typically limited to the second/third word of the sentence, the difference between both 'RealLabels' and 'AlterLabels' stress label sets lied approximately in one sixth of syllable stress labels, while the rest of the stress targets was identical.

The model used as a classifier was CART J48 decision tree trained on 'ctx2' features (left/right context with two syllables on each side) from 190 manually verified utterances from SPEECON dataset, so both training/testing data were read speech. Just to note, 'ctx2' features cause that each two initial and trailing syllables in the utterance are removed from the training/testing data themselves, but the information they provided is covered in the context. Also, removed syllable labels were not in the scope of real vs. alternative label differences and thus the ratio of individual different labels raised approximately from one sixth to one fourth, while the rest of the testing data labels was identical (and could not influence the computed scores difference).

Used feature set is the same as in the experiments described in section 6.6 consisting

generally of relative time distance between nuclei centers, pitch features, vowel normalized intensity features and vowel normalized spectral slope features.

The aim of experiment was to judge, if trained prosodic stress syllable model together with proposed scoring prosodic system of the utterance is able to distinguish correctly between real and alternative stress-group realizations or in other words, the situations, where examined prosodic scores $PFSC$ and $PFSC_w$ obtained for 'RealLabels' utterances stress labels were higher than scores for 'AlterLabels' syllabic stress labels. In fact, this task is direct application of N-best evaluation (where N=2) for ASR output without further ASR confidences.

Results

In the tables 8.2 and 8.3 are results obtained on the test set, which consisted of 130 couples of utterances composed from real and alternative footing segmentation using two score metrics PFSC and PFSCw described in section 8.1.1. Using comparison of those scores, it was possible to correctly identify the real utterance segmentation into stress-groups in 66.9% for PFSC score and in 73.1% for weighted PFSCw. The scores were equal in 15.4% of cases for PFSC and in 5.3% of cases for PFSCw respectively. Although error rate raised slightly from 17.7% for PFSC to 21.5% for PFSCw, this can be explained by the noticeable decrease of score equality situations ('cannot decide' lines) which marks expected narrowing of the "indecisiveness band" and is also accompanied by the success rate increase.

The column average score difference in both tables suggests, that the score differences which lead to 'Correct' result were higher than those for 'Wrong' decisions. This means that used model is more sure about 'Correct' results than about 'Wrong' ones, which is a good sign next to the raw accuracy obtained by the 'relative count' parameter. It also means that the ratio of correctly chosen variants compared to wrong decisions might be increased by setting a threshold of the score difference. When this score difference threshold would not be reached, the tested sample would come under 'Cannot decide' category, whose representation will grow.

For comparison of achieved results by machine learning methods with baseline of human perception, the listening test with two participants was accomplished on identical test set with identical assignment – to decide for given audio file with utterance, which variant out of two possibilities was realized. Both the participants were native Czech speakers without any phonetic education, but both were already familiar with a notion of stress-groups in Czech. Participants could went through 130 utterances in unlimited time (pauses were allowed and used in both individual tests) with the ability of unlimited

PFSC score results	absolute cnt.	relative cnt.	avg. score difference
Correct	87/130	66.9%	0.268736
Cannot decide	20/130	15.4%	0.0
Wrong	23/130	17.7%	-0.216522

Table 8.2: Results of PFSC scores comparison on 130 utterances with real and alternative stress-group segmentation suggestions. Result of comparison is 'correct' if PFSC of real stress group segmentation achieved higher score than its unrealized alternative, 'cannot decide' means sharp equality of scores and 'wrong' is the case of higher score of unrealized alternative.

PFSCw score results	absolute cnt.	relative cnt.	avg. score difference
Correct	95/130	73.1%	0.2403
Cannot decide	7/130	5.3%	0.0
Wrong	28/130	21.5%	-0.175

Table 8.3: Results of weighted PFSCw scores comparison on 130 utterances with real and alternative stress-group segmentation suggestions. Result of comparison is 'correct' if PFSCw of real stress group segmentation achieved higher score than its unrealized alternative, 'cannot decide' means sharp equality of scores and 'wrong' is the case of higher score of unrealized alternative.

number of utterance repetitions if needed. Result of the listening test can be found in the table 8.4 with comparison to the trained classifier using PFSCw. Very similar results for both the participants suggested possible problems in the reference data, if their answers matched against the whole testing dataset. This is why their responses were compared against each other with the resulting 20.8% inter-participant inconsistency (not present in the table). This is more than correct rate of both of them against the reference and thus the hypothesis about errors in reference was rejected and the results of listening tests are expected to be valid.

To conclude, the used methods of machine learning allowed in this experiment to mimic the human perception of speech segmentation into stress group for chosen subset of ambiguous Czech utterances with accuracy of 73% (while they still fail for 22% of cases and in 5% of cases the used methods cannot decide between the correct and alternative version). The results show interesting deployment possibilities even of this particular

	Correct [%]	Cannot decide [%]	Wrong [%]
Participant1	84.6	0.0	15.4
Participant2	83.9	0.0	16.1
PFSCw	73.1	5.3	21.5

Table 8.4: Result of the listening test on 130 ambiguous utterances in terms of stress-group segmentation. Comparison with results obtained by weighted prosodic utterance score PFSCw applied on the trained stress classifier is provided.

stress-group model for the real Czech ASR N-best hypothesis prosodic evaluation.

Chapter 9

Conclusions

In this chapter the work results in comparison with the set goals are being investigated. Three main areas regarding the research presented in this thesis are discussed.

9.1 Pitch Detection Algorithms

- several modifications to pitch detection algorithm evaluation criteria was suggested and implemented into the evaluation framework working on the top of three speech pitch reference databases
- various algorithms were compared using proposed evaluation framework with a focus on speech signals
- several modifications of existing algorithms were proposed, mainly dealing with the post-processing stage with Viterbi dynamic decoding (showing the importance of semitone distance measure as transition probability function between candidates, extending existing transition probability function by the concept of temporal forgetting)

9.2 Czech stress-group system

- a deep study was performed regarding the field of Czech stress-group system
- lexical tool for automatic Czech sentence division into stress groups was developed using phonetic rules
- prosodically enriched (in terms of lexical stress suitable for machine learning processing) Czech database based on Czech SPEECON was prepared, part of it was manually verified and labeled

- two more corpora originally used in phonetic research were adopted and prepared for automatic machine processing
- suitable acoustic feature set for Czech stress-group detection was chosen and system for its extraction was developed
- a CART decision tree classifier-based solution from acoustic data for Czech stress-group detection task was realized
- an objective measure for evaluating ASR hypothesis in terms of its prosodic probability was propose allowing ASR N-best assessment
- a usability of suggested measure was experimentally verified on the set of ambiguous utterances in terms of possible stress-group segmentation; proposed machine learning approach score was 73% while two participants of the listening test of the identical task achieved score of 85%
- a scheme for Czech ASR lattice rescoring using prosodic information about utterance stress-group segmentation was proposed

9.3 Czech phrase modality

- a database suitable for machine learning approach was prepared and manually annotated
- a MLP classifier-based solution from pure acoustic data for Czech phrase modality classification task was realized

9.4 Research and practical impacts

In this doctoral thesis mainly an approach towards human perception in the task of stress-group perception in Czech was achieved with practical utilization in ASR system. The final integration of the proposed algorithms and mechanisms into real Czech ASR system is a logical follow-up of presented thesis. Next to it, the field of Czech lexical stress and modality classification can be still considered as open research topic, although several findings presented in this work can extend actual knowledge in those areas.

Appendix A

Pitch Detection Algorithms evaluation

SPEECON CH0	VE	UE	GEH	GEL	GE	GEH10	GEL10	GE10	HE	DE	$\overline{\Delta}$	σ
PDA	[%]	[%]	[%]	[%]	[%]	[%]	[%]	[%]	[%]	[%]	[cents]	[cents]
Praat_{ac}	9.21	3.68	0.54	2.02	2.56	1.74	2.81	4.55	1.51	0.07	-9.19	221.66
Praat_{cc}	9.06	3.04	0.77	2.30	3.07	2.30	3.26	5.56	1.81	0.08	-13.37	225.28
$\text{Praat}_{\text{shs}}$	18.88	27.32	0.73	3.01	3.74	1.79	4.33	6.11	1.74	0.11	-35.50	236.03
$\text{Wavesurfer}_{\text{ESPS}}$	5.46	5.23	0.80	2.84	3.64	2.40	4.34	6.73	1.78	0.06	-26.81	226.31
$\text{MNFBC}_{\text{no\_post}}$	10.27	10.82	1.03	3.20	4.24	2.67	4.96	7.62	2.41	0.12	-39.89	288.34
$\text{MNFBC}_{\text{medfilt5}}$	8.85	11.33	0.59	3.09	3.69	2.27	5.27	7.54	2.12	0.04	-46.35	247.64
$\text{MNFBC}_{\text{Vit\_CentDiff}}$	10.27	10.82	0.80	1.79	2.59	2.49	3.64	6.13	1.17	0.09	-28.44	205.17
Kotnik2009	7.28	2.25	1.39	0.15	1.54	N/A	N/A	N/A	N/A	N/A	N/A	N/A

SPEECON CH0	Gross Errors in frequency bands, format (GE/GE10 @ (100-VE)) [%]				
PDA	<88 Hz	88 Hz - 141 Hz	141 Hz - 225 Hz	225 Hz – 353 Hz	>353 Hz
Praat_{ac}	2.5/7.4 @ 83.8	1.6/3.8 @ 89.2	3.2/4.9 @ 92.3	2.8/4.2 @ 92.5	25.4/35.7 @ 60.8
Praat_{cc}	4.4/8.4 @ 64.8	1.3/4.5 @ 91.8	4.1/6.1 @ 92.3	4.1/5.8 @ 91.5	30.1/39.4 @ 61.0
$\text{Praat}_{\text{shs}}$	4.2/10.2 @ 78.4	2.7/6.2 @ 81.1	4.2/5.7 @ 82.0	4.7/5.8 @ 78.9	34.9/40.2 @ 50.9
$\text{Wavesurfer}_{\text{ESPS}}$	3.8/9.9 @ 91.3	2.7/6.5 @ 94.6	4.1/6.6 @ 95.0	4.5/6.4 @ 93.8	35.9/40.6 @ 74.4
$\text{MNFBC}_{\text{no\_post}}$	4.4/8.0 @ 74.5	2.4/5.9 @ 87.9	5.4/8.6 @ 91.8	5.1/8.6 @ 92.7	37.1/39.4 @ 58.7
$\text{MNFBC}_{\text{medfilt5}}$	2.9/6.9 @ 76.3	1.9/6.0 @ 89.4	4.7/8.4 @ 92.7	4.8/8.7 @ 93.5	46.3/53.2 @ 61.6
$\text{MNFBC}_{\text{Vit\_CentDiff}}$	3.5/7.2 @ 74.5	1.6/5.2 @ 87.9	3.0/6.4 @ 91.8	3.5/7.3 @ 92.7	39.0/41.1 @ 58.7

Table A.1: SPEECON Channel 0 - overall results

SPEECON CH1	VE	UE	GEH	GEL	GE	GEH10	GEL10	GE10	HE	DE	$\overline{\Delta}$	σ
PDA	[%]	[%]	[%]	[%]	[%]	[%]	[%]	[%]	[%]	[%]	[cents]	[cents]
Praat_{ac}	20.04	11.07	13.05	3.42	16.47	14.15	4.09	18.24	2.34	1.87	250.79	837.33
Praat_{cc}	21.96	7.76	9.71	2.47	12.18	11.06	3.30	14.36	1.85	1.34	190.78	722.16
$\text{Praat}_{\text{shs}}$	22.10	30.87	1.37	5.86	7.22	2.37	7.18	9.55	3.09	0.34	-56.01	357.13
$\text{Wavesurfer}_{\text{ESPS}}$	25.49	6.82	2.24	4.18	6.43	3.67	5.39	9.06	2.67	0.35	-8.76	413.06
$\text{MNFBC}_{\text{no\_post}}$	24.49	9.28	14.67	3.87	18.54	16.17	5.28	21.45	2.78	2.05	257.66	879.18
$\text{MNFBC}_{\text{medfilt5}}$	23.26	9.34	13.51	3.57	17.08	15.12	5.34	20.46	2.47	1.96	235.07	843.43
$\text{MNFBC}_{\text{Vit\_CentDiff}}$	24.49	9.28	14.15	1.79	15.94	15.75	3.30	19.05	1.07	1.76	275.20	839.61
Kotnik2009	9.28	5.48	1.62	3.17	4.79	N/A	N/A	N/A	N/A	N/A	N/A	N/A

SPEECON CH1	Gross Errors in frequency bands, format (GE/GE10 @ (100-VE)) [%]				
PDA	<88 Hz	88 Hz - 141 Hz	141 Hz - 225 Hz	225 Hz – 353 Hz	>353 Hz
Praat_{ac}	36.1/39.2 @ 65.2	21.1/23 @ 76.5	13.7/15.4 @ 82.1	8.7/10.4 @ 87.7	31.3/39.4 @ 56.4
Praat_{cc}	34.9/37.4 @ 57.7	14.1/16.8 @ 75.6	10.6/12.4 @ 79.9	7.6/9.4 @ 85.4	28.6/37.1 @ 53.7
$\text{Praat}_{\text{shs}}$	10.0/16.8 @ 60.2	6.1/9.9 @ 75.6	7.6/8.9 @ 80.2	8.1/9.0 @ 81.9	49.6/52.8 @ 57.2
$\text{Wavesurfer}_{\text{ESPS}}$	6.6/12.3 @ 51.3	6.9/10.3 @ 71.9	6.2/8.4 @ 76.7	5.7/7.4 @ 82.1	32.6/36.0 @ 49.6
$\text{MNFBC}_{\text{no\_post}}$	41.1/43.1 @ 54.2	22.5/25.3 @ 72.1	16.2/19.1 @ 78.1	11.4/14.6 @ 83.5	44.1/48.1 @ 45.2
$\text{MNFBC}_{\text{medfilt5}}$	39.6/42.1 @ 55.4	21.1/24.5 @ 73.4	14.7/18 @ 79.3	10.0/13.6 @ 84.5	44.9/52.1 @ 46.2
$\text{MNFBC}_{\text{Vit\_CentDiff}}$	41.4/43.9 @ 54.2	21.6/24.6 @ 72.1	12.4/15.5 @ 78.1	8.1/11.7 @ 83.5	41.9/47.2 @ 45.2

Table A.2: SPEECON Channel 1 - overall results

KEELE DB	VE	UE	GEH	GEL	GE	GEH10	GEL10	GE10	HE	DE	$\overline{\Delta}$	σ
PDA	[%]	[%]	[%]	[%]	[%]	[%]	[%]	[%]	[%]	[%]	[cents]	[cents]
Praat_{ac}	9.32	2.88	0.73	0.57	1.30	1.34	0.72	2.07	0.56	0.25	7.90	205.47
Praat_{cc}	7.59	2.81	0.65	1.06	1.71	1.67	1.30	2.97	0.99	0.21	-1.95	191.54
$\text{Praat}_{\text{shs}}$	18.33	18.10	1.08	0.65	1.74	1.72	1.04	2.76	0.51	0.74	-6.37	163.37
$\text{Wavesurfer}_{\text{ESPS}}$	4.64	4.66	1.22	0.85	2.07	2.28	1.63	3.90	0.63	0.47	-1.99	185.06
$\text{MNFBC}_{\text{no\_post}}$	13.85	19.33	1.85	1.95	3.80	2.91	3.61	6.51	1.68	0.87	-19.85	292.53
$\text{MNFBC}_{\text{medfilt5}}$	12.69	17.61	1.28	1.53	2.81	2.24	3.69	5.93	1.21	0.64	-25.97	236.69
$\text{MNFBC}_{\text{Vit\_CentDiff}}$	13.85	19.33	1.91	0.46	2.37	3.07	2.11	5.18	0.25	1.01	-4.13	228.50

KEELE DB	Gross Errors in frequency bands, format (GE/GE10 @ (100-VE)) [%]				
PDA	<88 Hz	88 Hz - 141 Hz	141 Hz - 225 Hz	225 Hz – 353 Hz	>353 Hz
Praat_{ac}	3.2/6.1 @ 78.6	1.8/2.8 @ 87.0	1.1/1.5 @ 93.9	0.1/0.3 @ 97.5	0/0 @ 100
Praat_{cc}	2.4/5.4 @ 83.0	1.9/3.2 @ 90.6	2.2/3.1 @ 94.6	0.2/1.2 @ 96.3	0/0 @ 100
$\text{Praat}_{\text{shs}}$	7.4/10.2 @ 79.6	1.6/3.3 @ 80.9	1.2/1.6 @ 85.1	0.2/0.4 @ 77.2	0/0 @ 53.1
$\text{Wavesurfer}_{\text{ESPS}}$	4.2/7.7 @ 91.6	2.9/5.2 @ 95.5	1.7/3.1 @ 95.8	0.1/1.3 @ 96.2	0/0 @ 96.2
$\text{MNFBC}_{\text{no\_post}}$	14.1/17.7 @ 46.2	2.5/5.7 @ 86.3	4.7/6.8 @ 91.4	1.6/4.7 @ 96.0	0/0 @ 89.2
$\text{MNFBC}_{\text{medfilt5}}$	10.5/14.3 @ 46.7	2.0/6.1 @ 87.3	3.5/5.7 @ 93.2	0.8/3.8 @ 96.3	0/0 @ 89.2
$\text{MNFBC}_{\text{Vit\_CentDiff}}$	14.3/18.3 @ 46.2	3.0/6.1 @ 86.3	1.4/3.6 @ 91.4	0.3/3.4 @ 96.0	0/0 @ 89.2

Table A.3: KEELE DB - overall results

BAGSHAW DB	VE	UE	GEH	GEL	GE	GEH10	GEL10	GE10	HE	DE	$\overline{\Delta}$	σ
PDA	[%]	[%]	[%]	[%]	[%]	[%]	[%]	[%]	[%]	[%]	[cents]	[cents]
Praat_{ac}	7.48	4.98	0.67	0.46	1.13	2.05	0.96	3.01	0.29	0.16	10.67	140.13
Praat_{cc}	5.45	4.65	0.56	0.77	1.34	1.99	1.21	3.20	0.56	0.08	6.07	145.85
$\text{Praat}_{\text{shs}}$	19.16	20.33	0.49	0.70	1.19	1.80	1.36	3.16	0.32	0.11	-6.24	141.99
$\text{Wavesurfer}_{\text{ESPS}}$	4.37	6.47	0.79	1.48	2.27	1.87	2.31	4.18	0.87	0.17	-6.25	170.15
$\text{MNFBC}_{\text{no\_post}}$	10.76	18.46	0.96	2.18	3.13	2.58	4.41	6.99	1.33	0.19	-34.01	216.10
$\text{MNFBC}_{\text{medfilt5}}$	9.44	17.95	0.86	1.73	2.59	2.43	4.30	6.73	0.73	0.15	-31.59	185.45
$\text{MNFBC}_{\text{Vit\_CentDiff}}$	10.76	18.46	0.98	1.25	2.23	2.63	3.52	6.15	0.53	0.22	-23.31	162.80

BAGSHAW DB	Gross Errors in frequency bands, format (GE/GE10 @ (100-VE)) [%]				
PDA	<88 Hz	88 Hz - 141 Hz	141 Hz - 225 Hz	225 Hz – 353 Hz	>353 Hz
Praat_{ac}	3.3/5.4 @ 64.6	0.4/1.5 @ 90.0	2.7/5.7 @ 94.0	0.8/2.7 @ 95.3	15.7/39.8 @ 70.3
Praat_{cc}	4.1/6.6 @ 75.7	0.7/1.7 @ 95.2	3.2/6.4 @ 95.2	0.8/2.7 @ 94.7	14.8/42.0 @ 68.6
$\text{Praat}_{\text{shs}}$	5.5/8.7 @ 78.9	0.8/2.1 @ 83.3	1.9/4.8 @ 81.4	0.9/2.9 @ 78.6	21.1/49.3 @ 60.2
$\text{Wavesurfer}_{\text{ESPS}}$	12.6/16.5 @ 87.5	1.6/3.2 @ 96.3	3.3/6.0 @ 96.2	1.8/3.5 @ 95.1	25.5/50.0 @ 83.1
$\text{MNFBC}_{\text{no\_post}}$	11.0/18.2 @ 64.4	1.6/6.7 @ 86.6	5.7/9.5 @ 89.8	2.7/5.5 @ 92.4	41.3/58.7 @ 78.0
$\text{MNFBC}_{\text{medfilt5}}$	9.7/14.4 @ 65.0	1.9/7.4 @ 88.7	4.3/8.2 @ 91.1	2.0/5.1 @ 93.1	43.0/59.1 @ 78.8
$\text{MNFBC}_{\text{Vit\_CentDiff}}$	11.3/18.5 @ 64.4	1.6/6.7 @ 86.6	4.4/8.2 @ 89.8	1.3/4.2 @ 92.4	38.0/58.7 @ 78.0

Table A.4: CSTR BAGSHAW DB - overall results

Appendix B

Example of real PDA halving errors

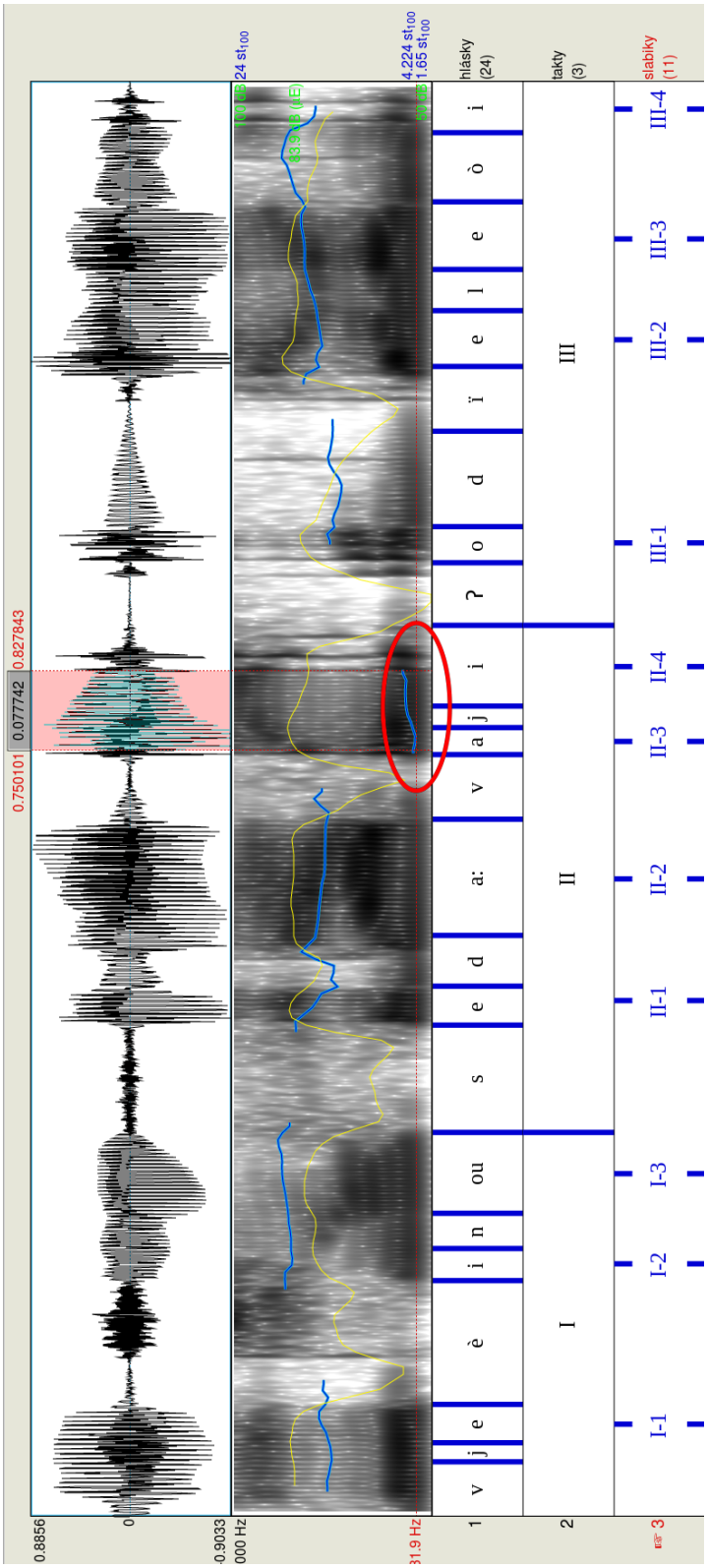


Figure B.1: Example of real halving octave errors represented by series of two syllables in a row (in the red circle) estimated by Praat [112] for utterance *OPE_pm23* from Helena Spilková diploma thesis corpus. Pitch contour is marked by blue line (units semitones related to 100 Hz), while intensity being yellow contour. Proposed octave errors resolving algorithm 6.3.5 is able to detect and recover from even this 2-syllable series.

Appendix C

List of ambiguous utterances from Helena Spilková diploma thesis

Variant "x"	**Variant "y"**
překvapil hobojový mistr	překvapil ho bojový mistr
ukázal mizející otvor	ukázal mi zející otvor
rodiče jídávali málo	rodiče jí dávali málo
většinou sedávají odděleně	většinou se dávají odděleně
pánové sestrojili podle	pánové se strojili podle
obraz bezvýrazně zbarvených	obraz bez výrazně zbarvených
říkat bezvýznamně znějící	říkat bez významně znějící
letěl nadstandardně vybaveným	letěl nad standardně vybaveným
nebyl podprůměrně vzrostlým	nebyl pod průměrně vzrostlým
zahájil předminule slíbeným	zahájil před minule slíbeným
potkala ilegálního přistěhovalce	potkala i legálního přistěhovalce
opominula irelevantní údaje	opominula i relevantní údaje
hledal půlměsíce v mešitě	hledal půl měsíce v mešitě
majitele tříbarevných papoušků	majitele tří barevných papoušků
obsahu třílitrových soudků	obsahu tří litrových soudků
výrobce dvoumetrových mečů	výrobce dvou metrových mečů
tato převelice zajímavá	tato pře velice zajímavá
snědl čtvrtkilového kraba	snědl čtvrt kilového kraba
cílem čtyřhodinových pořadů	cílem čtyř hodinových pořadů

Table C.1: List of ambiguous utterances from Helena Spilková diploma thesis

Bibliography

[1] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer handbook of speech processing*. Springer Science & Business Media, 2007.

[2] C. Starr, C. Evers, and L. Starr, *Biology: A human emphasis*, ser. Cengage Advantage Books. Cengage Learning; 8 edition, 2010, ISBN: 0538757027.

[3] Z. Otčenášek, *O subjektivním hodnocení zvuku*, 1st. Akademie múzických umění v Praze, 2008, ISBN: 978-80-7331-113-1.

[4] J. F. Ransome and J. Rittinghouse, *Voice over internet protocol (VoIP) security*. Digital Press, 2005, p. 432.

[5] J. UHLÍŘ, *Technologie hlasových komunikací*. ČVUT Praha, 2007, ISBN: 978-80-01-03888-8.

[6] M. Beckman, *Stress and non-stress accent*. Dordrecht: Foris public, 1986.

[7] D. Hermes, "Stylization of pitch contours", in *Methods in Empirical Prosody Research*, Berlin, New York: De Gruyter, 2006, pp. 29–62.

[8] V. Syrový, *Hudební akustika*, 2nd. HAMU Praha, 2008, ISBN: 978-80-7331-127-8.

[9] H. Fletcher and W. Munson, "Loudness, its definition, measurement and calculation", *Journal of the Acoustic Society of America*, vol. 5, pp. 82–108, 1933.

[10] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[11] J. Martens, "Deep learning via Hessian-free optimization", in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 735–742.

[12] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets", *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[13] M. J. Gales and P. C. Woodland, "Mean and variance adaptation within the mllr framework", *Computer Speech & Language*, vol. 10, no. 4, pp. 249–264, 1996.

[14] G. Saon, T. Sercu, S. Rennie, and H.-K. J. Kuo, "The IBM 2016 english conversational telephone speech recognition system", *ArXiv e-prints*, Apr. 2016. arXiv: 1604.08242.

[15] S. Goldwater, D. Jurafsky, and C. D. Manning, "Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates", *Speech Communication*, vol. 52, no. 3, pp. 181–200, 2010.

[16] P. A. Luce and D. B. Pisoni, "Recognizing spoken words: The neighborhood activation model", *Ear and hearing*, vol. 19, no. 1, pp. 1–36, 1998.

[17] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition", in *Proceedings of INTERSPEECH*, 2015.

[18] F. SANTIAGO, M ADDA-DECKER, and C. DUTREY, "Towards a typology of ASR errors via syntax-prosody mapping", in *Errors by Humans and Machines in multimedia, multimodal and multilingual data processing (ERRARE'15 Workshop)*, Sinaia, Romania, 2015.

[19] M. Boháč, J. Nouza, and K. Blavka, "Investigation on most frequent errors in large-scale speech recognition applications", in *Proc. of International Conference on Text, Speech and Dialogue*, Springer, 2012, pp. 520–527.

[20] Z. Palková, "The set of phonetic rules as a basis for the prosodic component of an automatic TTS synthesis in Czech", *Phonetica Pragensia*, vol. 10, pp. 33–46, 2004.

[21] M. Wester, C. Valentini-Botinhao, and G. E. Henter, "Are we using enough listeners? No! An empirically-supported critique of Interspeech 2014 TTS evaluations", in *Proc. of Interspeech*, 2015.

[22] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, "Prosody contour prediction with long short-term memory (LSTM), bi-directional, deep recurrent neural networks", in *Proceedings of InterSpeech 2014*, 2014, pp. 2268–2272.

[23] V. Sethu, E. Ambikairajah, and J. Epps, "On the use of speech parameter contours for emotion recognition", *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, pp. 1–14, 2013.

[24] J. Přibil and A. Přibilová, "Evaluation of influence of spectral and prosodic features on GMM classification of Czech and Slovak emotional speech", *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, pp. 1–22, 2013.

[25] M. Kockmann, "Subspace modeling of prosodic features for speaker verification", PhD thesis, Brno, CZ, 2012, p. 122.

[26] *Fonetická identifikace mluvčího*. Filozofická fakulta UK, Praha, 2014, ISBN: 978-80-7308-548-3.

[27] D. Martínez, L. Burget, L. Ferrer, and N. Scheffer, "iVector-based prosodic system for language identification", in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2012*, IEEE, 2012, pp. 4861–4864.

[28] N. Singh, *Prosodic featured based automatic language identification*. Educreation Publishing, 2015, ISBN: 9385247050.

[29] R. Razo, A. Lozano-Diez, J. Gonzalez-Dominguez, D. T. Toledano, and J. Gonzalez-Rodriguez, "Language identification in short utterances using long short-term memory (LSTM) recurrent neural networks", *PloS one*, vol. 11, no. 1, 2016.

[30] N. Veilleux and M. Ostendorf, "Prosody/parse scoring and its application in ATIS", in *Proceedings of the ARPA Workshop on Human Language Technology*, 1993, pp. 335–340.

[31] W. Wahlster, *Verbmobil: Foundations of speech-to-speech translation*, ser. Artificial intelligence. Springer, 2000, ISBN: 9783540677833. [Online]. Available: `http://books.google.com/books?id=RiT0aAzeudkC`.

[32] F. Gallwitz, H. Niemann, E. Nöth, and V. Warnke, "Integrated recognition of words and prosodic phrase boundaries.", *Speech Communication*, pp. 81–95, 2002.

[33] R. Kompe, A. Kießling, H. Niemann, E. Nöth, E. G. Schukat-Talamazzini, A. Zottmann, and A. Batliner, "Prosodic scoring of word hypotheses graphs", in *EUROSPEECH'95*, 1995.

[34] K. Hirose, N. Minematsu, Y. Hashimoto, and K. Iwano, "Continuous speech recognition of Japanese using prosodic word boundaries detected by mora transition modeling of fundamental frequency contours", in *In Proceedings of the Workshop on Prosody in Automatic Speech Recognition and Understanding*, 2001, pp. 61–66.

[35] D. Beeferman, A. Berger, and J. Lafferty, "CYBERPUNC: A lightweight punctuation annotation system for speech", in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1998, pp. 689–692.

[36] W. Lu and H. T. Ng, "Better punctuation prediction with dynamic conditional random fields", in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '10, Cambridge, Massachusetts: Association for Computational Linguistics, 2010, pp. 177–186.

[37] C. J. Chen, "Speech recognition with automatic punctuation", in *Proc. of EUROSPEECH'99*, 1999, pp. 447–450. [Online]. Available: `http://www.isca-speech.org/archive/eurospeech\_1999/e99\_0447.html`.

[38] E. Shriberg, R. Bates, A. Stolcke, P. Taylor, A. Erringer, M. Gregory, L. Heintzelman, T. Metzler, A. Oduro, and T. The, "Can prosody aid the automatic classification of dialog acts in conversational speech?", vol. 41, no. 3-4, pp. 439–487, 1998.

[39] J. hwan Kim and P. C. Woodland, "The use of prosody in a combined system for punctuation generation and speech recognition", in *Proc. of EUROSPEECH*, 2001, pp. 2757–2760.

[40] H. Christensen, Y. Gotoh, and S. Renals, "Punctuation annotation using statistical prosody models", in *Proc. of ISCA Workshop on Prosody in Speech Recognition and Understanding*, Red Bank, NJ, USA, 2001, pp. 35–40.

[41] J. Huang and G. Zweig, "Maximum entropy model for punctuation annotation from speech", in *Proc. International Conference on Spoken Language Processing (ICSLP)*, 2002, pp. 917–920.

[42] V. Strom, "Detection of accents, phrase boundaries and sentence modality in German with prosodic features", in *EUROSPEECH*, vol. 3, 1995, pp. 2039–2042.

[43] P. Král, J. Klečková, and C. Cerisara, "Sentence modality recognition in French based on prosody", *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, vol. 1, no. 8, pp. 2552–2555, 2007.

[44] Y. Gotoh and S. Renals, "Sentence boundary detection in broadcast speech transcripts", in *Proc. of ISCA Workshop: Automatic Speech Recognition: Challenges for the new Millennium ASR-2000*, 2000, pp. 228–235.

[45] S. Ananthakrishnan and S. Narayanan, "Unsupervised adaptation of categorical prosody models for prosody labeling and speech recognition", *IEEE Trans. Audio, Speech and Language Processing*, vol. 17, no. 1, pp. 138–149, 2009. [Online]. Available: `http://sail.usc.edu/publications/Shankar-TASLP2009.pdf`.

[46] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, "The Boston University Radio News Corpus", Boston University, Tech. Rep., Mar. 1995. [Online]. Available: `http://ssli.ee.washington.edu/papers/radionews-tech.ps`.

[47] S.-H. Chen, J.-H. Yang, C.-Y. Chiang, M.-C. Liu, and Y.-R. Wang, "A new prosody-assisted mandarin asr system", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1669–1684, 2012.

[48] K. Vicsi and G. Szaszák, "Using prosody to improve automatic speech recognition", *Speech Communication*, vol. 52, no. 5, pp. 413–426, 2010, ISSN: 0167-6393. DOI: http://dx.doi.org/10.1016/j.specom.2010.01.003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639310000129.

[49] K. Vicsi and G. Szaszak, "Automatic segmentation of continuous speech on word level based on supra-segmental features", *International Journal of Speech Technology*, vol. 8, pp. 363–370, 4 2005, 10.1007/s10772-006-8534-z, ISSN: 1381-2416. [Online]. Available: http://dx.doi.org/10.1007/s10772-006-8534-z.

[50] K. Silverman, M. Beckman, J. Pierrehumbert, M. Ostendorf, C. Wightman, P. Price, and J. Hirschberg, "TOBI: A standard scheme for labeling prosody", in *International Conference on Spoken Language Processing*, 1992.

[51] P. N. Garner, R. Clark, J.-P. Goldman, P.-E. Honnet, M. Ivanova, A. Lazaridis, H. Liang, B. Pfister, M. S. Ribeiro, E. Wehrli, and J. Yamagishi, "Translation and prosody in Swiss languages", in *Nouveaux cahiers de linguistique francaise*, 2014.

[52] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue", *Computer Speech and Language*, vol. 25, no. 3, pp. 601–634, Jul. 2011, ISSN: 0885-2308.

[53] S. Oviatt, C. Darves, R. Coulston, and M. Wesson, "Speech convergence with animated personas", in *Spoken Multimodal Human-Computer Dialogue in Mobile Environments*, W. Minker, D. Bühler, and L. Dybkjær, Eds. Dordrecht: Springer Netherlands, 2005, pp. 379–397, ISBN: 978-1-4020-3075-8.

[54] J. Kolář, E. Shriberg, and Y. Liu, "Using prosody for automatic sentence segmentation of multi-party meetings", in *Proc. of International Conference on Text, Speech and Dialogue*, Springer, 2006, pp. 629–636.

[55] J. Kolář, "Automatic segmentation of speech into sentence-like units", PhD thesis, University of West Bohemia in Pilsen, 2008.

[56] J. Kolář, Y. Liu, and E. Shriberg, "Genre effects on automatic sentence segmentation of speech: A comparison of broadcast news and broadcast conversations", in *Proc. of 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2009, pp. 4701–4704.

[57] G. Szaszák and A. Beke, "Exploiting prosody for syntactic analysis in automatic speech understanding", *Journal of Language Modelling*, no. 1, pp. 143–172, 2012.

[58] J. Kolář and Y. Liu, "Automatic sentence boundary detection in conversational speech: A cross-lingual evaluation on English and Czech", in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 5258–5261.

[59] M. Breen, L. C. Dilley, J. Kraemer, and E. Gibson, "Inter-transcriber reliability for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch)", *Corpus Linguist. Ling.*, vol. 8, no. 2, pp. 277–312, 2012.

[60] J. Vaissière, "Perception of intonation", in *The Handbook of Speech Perception*, Oxford, UK: Blackwell Publishing Ltd, 2005.

[61] "Global and local evaluation of prosody: Discrete target and just noticeable differences", in *Proc. of Speech Prosody 2008*, Campinas: University of Campinas, 2008, pp. 727–730.

[62] A. Belotel-Grenié and M. Grenié, "The creaky voice phonation and the organisation of Chinese discourse", in *International symposium on tonal aspects of languages: With emphasis on tone languages*, 2004.

[63] T. Duběda, *Jazyky a jejich zvuky. univerzálie a typologie ve fonetice a fonologii*. Karolinum, Praha, 2005, ISBN: 8024610736.

[64] A. M. Sluijter and V. J. Van Heuven, "Spectral balance as an acoustic correlate of linguistic stress", *The Journal of the Acoustical society of America*, vol. 100, no. 4, pp. 2471–2485, 1996.

[65] R. Skarnitzl, P. Šturm, and J. Volín, *Zvuková báze řečové komunikace: Fonetický a fonologický popis řeči*. Karolinum: Praha, 2016, ISBN: 978-80-246-3300-8 (PDF).

[66] L. Weingartová and J. Volín, "Short-term spectral slope measures and their sensitivity to speaker, vowel identity and prominence", *Akustické listy*, vol. 20, no. 1, pp. 5–12, 2014.

[67] T. Duběda, "K izosylabičnosti a izochronnosti v češtině [on syllable-timing and stress-timing in Czech]", in *Sborník z Konference česko-slovenské pobočky ISPhS 2004*, Original document in Czech, Univerzita Karlova v Praze, Filozofická fakulta, 2004, pp. 19–28.

[68] J. Dankovičová and V. Dellwo, "Czech speech rhythm and the rhythm class hypothesis", in *Proc. 16th ICPhS*, 2007, pp. 1241–1244.

[69] E. Grabe and E. Low, "Durational variability in speech and the rhythm class hypothesis", in *Papers in Laboratory Phonology 7, Berlin, New York: Mouton de Gruyter*, C. Gussenhoven and N. Warner, Eds., 2004, pp. 62–67.

[70] F. Ramus, M. Nespor, and J. Mehler, "Correlates of linguistic rhythm in the speech signal", *Cognition*, vol. 73, pp. 265–292, 1999, ISSN: 1210-2709.

[71] P. Janota, "An experiment concerning the perception of stress by czech listeners", *Phonetica Pragensia II*, pp. 45–68, 1967.

[72] Z. Palková, "Einige Beziehungen zwischen prosodischen Merkmalen im Tschechischen", in *Proceedings of the Fourteenth International Congress of Linguists*, Berlin, 1987, pp. 507–510.

[73] T. Duběda and J. Raab, "Pitch accents, boundary tones and contours: Automatic learning of Czech intonation", in *International Conference on Text, Speech and Dialogue*, Springer, 2008, pp. 293–301.

[74] Z. Palková, *Fonetika a fonologie češtiny [phonetics and phonology of Czech]*. Karolinum, Praha, 1994, ISBN: 80-7066-843-1.

[75] Z. Palková and J. Volín, "The role of F0 contours in determining foot boundaries in Czech", in *Proceedings of the 15th ICPhS, Barcelona*, vol. 2, 2003, pp. 1783–1786, ISBN: 1-876346-49-3.

[76] H. Spilková, "Intonace jednoslabičného slova jako součásti víceslabičného taktu [in Czech]", Master's thesis, Praha: Fonetický ústav Filozofické fakulty Univerzity Karlovy, not published, 2007.

[77] J. Volín, "Z intonace čtených zpravodajství: Výška první slabiky v taktu", *Čeština doma a ve světě*, vol. 1-2, pp. 89–96, 2008, ISSN: 1210-9339.

[78] Z. Palková, J. Veroňková, J. Volín, and R. Skarnitzl, "Stabilizace některých termínů pro fonetický popis češtiny v závislosti na nových výsledcích výzkumu", in *Sborník z Konference česko-slovenské pobočky ISPhS*, 2004, pp. 65–74.

[79] R. Skarnitzl, "O slovním přízvuku na jednoslabičných předložkách v češtině", *Naše řeč*, vol. 97, pp. 78–91,

[80] T. Duběda, ""Flat pitch accents" in Czech", in *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, 2010, pp. 1756–1759.

[81] T. Duběda, "Towards an inventory of pitch accents for read Czech", *Slovo a slovesnost*, vol. 71, no. 1, pp. 3–13, 2011.

[82] J. Chamonikolasová, *Intonation in english and czech dialogues*. Masarykova univerzita: Brno, 2007, ISBN: 978-80-210-4468-8.

[83] Z. Palková *et al.*, "Melodemes in speech: Their stability and confusions", *Acta Universitatis Carolinae Philologica*, no. 1, pp. 55–68, 2014.

[84] T. Duběda *et al.*, "Czech intonation: A tonal approach", *Slovo a slovesnost*, vol. 75, no. 2, pp. 83–98, 2014.

[85] D. Hirst and A. Cristo, *Intonation systems: A survey of twenty languages*. Cambridge University Press, 1998, ISBN: 9780521395502. [Online]. Available: `http://books.google.com/books?id=LClvNiI4k0sC`.

[86] E. Grabe, B. Post, and F. Nolan, "Modelling intonational variation in english: The ivie system", in *Proceedings of Prosody*, vol. 5, 2000, pp. 1–57.

[87] B. Post, E. Delais-Roussarie, and A.-C. Simon, "IVTS, un système de transcription pour la variation prosodique", in *Bulletin PFC*, vol. 6, 2006, pp. 51–68.

[88] L. C. Dilley and M. Brown, "The RaP (Rhythm and Pitch) Labeling System", *Unpublished manuscript*, 2005.

[89] M. Breen, L. C. Dilley, J. Kraemer, and E. Gibson, "Inter-transcriber reliability for two systems of prosodic annotation: Tobi (tones and break indices) and rap (rhythm and pitch)", 2012.

[90] A. K. Syrdal, J. Hirschberg, J. McGory, and M. E. Beckman, "Automatic ToBI prediction and alignment to speed manual labeling of prosody.", *Speech Communication*, pp. 135–151, 2001.

[91] M. Rusko, R. Sabo, and M. Dzúr, "Sk-ToBI scheme for phonological prosody annotation in Slovak", in *Proc. of TSD 2007*, V. Matousek and P. Mautner, Eds., ser. Lecture Notes in Computer Science, vol. 4629, Springer, Aug. 27, 2007, pp. 334–341, ISBN: 978-3-540-74627-0. [Online]. Available: `http://dblp.uni-trier.de/db/conf/tsd/tsd2007.html#RuskoSD07`.

[92] C. W. Wightman, "ToBI Or Not ToBI?", in *Proc. of Speech Prosody 2002*, 2002, pp. 25–29. [Online]. Available: `http://www.isca-speech.org/archive/sp2002/sp02\_025.pdf`.

[93] P. C. Bagshaw, "An investigation of acoustic events related to sentential stress and pitch accents, in English", *Speech Communication*, vol. 13, no. 3, pp. 333–342, 1993.

[94] K. M. Higgins, *The music between us: Is music a universal language?* University of Chicago Press, 2012.

[95] J. Steele, "An essay towards establishing the melody and measure of speech to be expressed and perpetuated by peculiar symbol", 1775.

[96] H. Hatfield, "Joshua Steele 1775: Speech intonation and music tonality", University of Hawai'i Manoa, 2010.

[97] D. A. Schwartz, C. Q. Howe, and D. Purves, "The statistical structure of human speech sounds predicts musical universals", *The Journal of Neuroscience*, vol. 23, no. 18, pp. 7160–7168, 2003.

[98] D. Ross, J. Choi, and D. Purves, "Musical intervals in speech", *Proceedings of the National Academy of Sciences*, vol. 104, no. 23, pp. 9852–9857, 2007.

[99] A. D. Patel, J. R. Iversen, and J. C. Rosenberg, "Comparing the rhythm and melody of speech and music: The case of british english and french", *The Journal of the Acoustical Society of America*, vol. 119, no. 5, pp. 3034–3047, 2006.

[100] A. D. Patel, *Music, language, and the brain*. New York: Oxford University Press, 2008, ISBN: 978-0-19-512375-3.

[101] K. Abdullah-Al-Mamun, F. Sarker, and G. Muhammad, "A high resolution pitch detection algorithm based on AMDF and ACF", *Journal of Scientific Research*, vol. 1, no. 3, 2009, ISSN: 2070-0245. [Online]. Available: `http://www.banglajol.info/index.php/JSR/article/view/2569`.

[102] D. Talkin, "A robust algorithm for pitch tracking (RAPT)", *Speech Coding and Synthesis, Elsevier Science*, pp. 495–518, 1995.

[103] H. Bořil and P. Pollák, "Direct time domain fundamental frequency estimation of speech in noisy conditions", *Proceedings of EUSIPCO 2004 (European Signal Processing Conference, Vol. 1)*, pp. 1003–1006, 2004.

[104] J. Bartošek, "Pitch detection algorithm evaluation framework", English, in *Proc. of 20th Czech-German Workshop on Speech Processing*, Prague: Institute of Photonics and Electronics AS CR, 2010, pp. 118–123, ISBN: 978-80-86269-21-4.

[105] C. B. F. Signol and J.-S. Lienard, "Evaluation of the pitch estimation algorithms in the monopitch and multipitch cases", in *Proceedings of Acoustics'08 Paris, France*, vol. 123, 2008.

[106] B. Kotnik, H. Höge, and Z. Kacic, "Evaluation of pitch detection algorithms in adverse conditions", *Proc. 3rd International Conference on Speech Prosody, Dresden, Germany*, pp. 149–152, 2006.

[107] G. M. F. Plante and A. Ainsworth, "A pitch extraction reference database", in *Proc. of Eurospeech'95*, 1995, pp. 837–840.

[108] B. Hiller, S. M. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching", in *Proceedings of the 3rd European Conference on Speech Communication and Technology*, 1993, pp. 1003–1006.

[109] M. Vondrášek and P. Pollák, "Methods for Speech SNR estimation: Evaluation Tool and Analysis of VAD Dependency", in *Radioengineering*, vol. 14, 2005.

[110] K. Sjölander and J. Beskow, "Wavesurfer - an open source speech tool", in *Sixth International Conference on Spoken Language Processing, ICSLP 2000 / INTERSPEECH 2000, Beijing, China, October 16-20, 2000*, vol. 4, ISCA, 2000, pp. 464–467. DOI: http://www.isca-speech.org/archive/icslp_2000/i00_4464.html.

[111] K. Sjölander. (1997-2004). The snack sound toolkit, [Online]. Available: http://www.speech.kth.se/snack/.

[112] P. Boersma, "Praat, a system for doing phonetics by computer.", *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[113] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm", *Information Theory, IEEE Transactions on*, vol. 13, no. 2, pp. 260–269, 1967, ISSN: 0018-9448. DOI: 10.1109/TIT.1967.1054010.

[114] H. Quast, O. Schreiner, and M. Schroeder, "Robust pitch tracking in the car environment", in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., IEEE International Conference on*, vol. 1, 2002, p. I.

[115] B. Kotnik, H. Höge, and Z. Kacic, "Noise robust F0 determination and epochmarking algorithms", *Signal Processing*, vol. 89, no. 12, pp. 2555–2569, 2009, ISSN: 0165-1684. DOI: DOI:10.1016/j.sigpro.2009.04.017.

[116] D. J. Hermes, "Measurement of pitch by subharmonic summation", *The Journal of the Acoustical Society of America*, vol. 83, no. 1, pp. 257–264, 1988. DOI: 10.1121/1.396427. [Online]. Available: http://link.aip.org/link/?JAS/83/257/1.

[117] J. Bartošek, "Pitch detection algorithm evaluation framework", English, in *Proc. of 20th Czech–German Workshop on Speech Processing*, Prague, CZ, 2010, pp. 118–123, ISBN: 978-80-86269-21-4.

[118] E. Churaňová, "The consonantal–vocalic structure of the Czech word and stress group", *AUC Philologica 1/2014, Phonetica Pragensia XIII*, vol. 1, pp. 79–90, 2014.

[119] ——, "The consonantal–vocalic structure of the Czech word and stress group", *AUC Philologica 1/2014, Phonetica Pragensia XIII*, vol. 1, pp. 79–90, 2014.

[120] T. Duběda and J. Votrubec, "Acoustic correlates of stress in Czech: A stochastic view", in *Proc. of 14th Cz-Ge Workshop*, 2004.

[121] ——, "Acoustic analysis of Czech stress: Intonation, duration and intensity revisited", in *Proc. of INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology*, Lisbon, Portugal, 2005, pp. 1429–1432.

[122] T. Duběda, "Intensity as a macroprosodic variable in Czech", in *Proc. of Speech Prosody 2006*, Dresden, 2006, pp. 185–188.

[123] L. C. Pols *et al.*, "Flexible, robust and efficient human speech processing versus present-day speech technology.", 1999.

[124] M. Kroul, "Automatic detection of emphasized words for performance enhancement of a Czech ASR system", in *Proceedings of 13th International Conference Speech and Computer (Specom 2009), St. Petersburg, Russia*, 2009, pp. 470–473, ISBN: 978-5-8088-0442-5.

[125] J. Bartošek, "Czech Text-to-Foot converter", in *Proc. of POSTER 2013 - 17th International Student Conference on Electrical Engineering*, Prague, 2013, ISBN: 978-80-01-05242-6.

[126] P. Hauser, *Základy skladby češtiny*. Brno: Masarykova univerzita, 2003, ISBN: 80-210-3113-1.

[127] C. Konrat and M. Jessen, "Fundamental frequency analysis: A collaborative exercise", in *Proc. of Conference of the International Association of Forensic Phonetics and Acoustics, Tampa, Florida, USA*, 2013.

[128] J. Volín, "Data volume requirements for reliable F0 normalization", in *Proc. of 17th Czech-German Workshop - Speech Processing*, R. Vích, Ed., Prague: Institute of Photonics and Electronics AS CR, 2007, pp. 62–67, ISBN: 978-80-86269-00-9.

[129] J. Lindh and A. Eriksson, "Robustness of long-time measures of fundamental frequency", in *Proceedings of Interspeech 2007*, Original document in Czech, Antwerp: ISCA, 2007, pp. 2025–2028.

[130] J. Volín, K. Poesová, and L. Weingartová, "Speech Melody Properties in English, Czech and Czech English: Reference and Interference", *Research in Language*, vol. 13, no. 1, pp. 107–123, 2015.

[131] J. Bartošek and V. Hanžl, "Foot detection in Czech using pitch information and HMM", in *Proc. of Text, Speech, and Dialogue 2013*, ser. Lecture Notes in Computer Science, Berlin: Springer, 2013, pp. 272–279, ISBN: 978-3-642-40584-6.

[132] S. Young and S. Young, "The HTK Hidden Markov Model toolkit: Design and philosophy", *Entropic Cambridge Research Laboratory, Ltd*, vol. 2, pp. 2–44, 1994.

[133] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006.

[134] C. D. Doran, "Incorporating punctuation into the sentence grammar: A lexicalized tree adjoining grammar perspective", PhD thesis, Faculties of the University of Pennsylvania, 1998.

[135] O. Tilk and T. Alumäe, "LSTM for punctuation restoration in speech transcripts", in *Interspeech 2015*, Dresden, Germany, 2015.

[136] E. Cho, C. Fügen, T. Herrmann, K. Kilgour, M. Mediani, C. Mohr, J. Niehues, K. Rottmann, C. Saam, S. Stüker, *et al.*, "A real-world system for simultaneous translation of german lectures", in *Proc. of INTERSPEECH*, 2013, pp. 3473–3477.

[137] E. Cho, J. Niehues, K. Kilgour, and A. Waibel, "Punctuation insertion for real-time spoken language translation", in *The International Workshop on Spoken Language Translation*, 2015.

[138] A. Pražák, Z. Loose, J. Trmal, J. V. Psutka, and J. Psutka, "Novel approach to live captioning through re-speaking: Tailoring speech recognition to re-speaker's needs.", in *Proc. of INTERSPEECH 2012*, 2012, pp. 1372–1375.

[139] J. Kleckova and V. Matousek, "Using prosodic characteristics in czech dialog system", 1997.

[140] J. Klečková, "Storing prosody attributes of spontaneous speech", in *Proc. of International Workshop on Text, Speech and Dialogue*, Springer, 1999, pp. 268–273.

[141] J. Kolář, J. Švec, and J. Psutka, "Automatic punctuation annotation in Czech broadcast news speech", *Proc. of SPECOM´ 2004*, 2004.

[142] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, "Prosody-based automatic segmentation of speech into sentences and topics", *Speech communication*, vol. 32, no. 1, pp. 127–154, 2000.

[143] J. Kolorenč, "Tvorba a adaptace lingvistické vrstvy pro systém rozpoznávání mluvené češtiny", PhD thesis, Liberec: Technická univerzita v Liberci, 2007.

[144] M. Boháč and K. Blavka, "Using suprasegmental information in recognized speech punctuation completion", in *Proc. of International Conference on Text, Speech, and Dialogue*, Springer, 2014, pp. 555–562.

[145] J. Bartošek and V. Hanžl, "Intonation based sentence modality classifier for Czech using artificial neural network", English, in *Advances in Nonlinear Speech Processing*, ser. Lecture Notes in Computer Science, Heidelberg: Springer, 2011, pp. 162–169, ISBN: 978-3-642-25019-4.

[146] L. Harada, "Complex temporal patterns detection over continuous data streams", in *Proceedings of the 6th East European Conference on Advances in Databases and Information Systems (ADBIS)*, Y. Manolopoulos and P. Návrat, Eds., ser. Lecture Notes in Computer Science, vol. 2435, London, UK: Springer-Verlag, 2002, pp. 401–414, ISBN: 3-540-44138-7.

[147] T. Jiang, Y. Feng, and B. Zhang, "Online detecting and predicting special patterns over financial data streams", *Journal of Universal Computer Science*, vol. 15, no. 13, pp. 2566–2585, 2009.

[148] J. L. Elman, "Finding structure in time", *Cognitive Science*, vol. 14, no. 2, pp. 179–211, 1990.

[149] G. Dorffner, "Neural networks for time series processing", *Neural Network World*, vol. 6, pp. 447–468, 1996.

[150] E. Haselsteiner and G. Pfurtscheller, "Using time-dependent neural networks for EEG classification", *IEEE Transactions on Rehabilitation Engineering*, vol. 8, pp. 457–463, 4 2000, ISSN: 1063-6528. DOI: 10.1109/86.895948.

[151] B. Zhou and J. Hu, "A dynamic pattern recognition approach based on neural network for stock time-series", in *Proc. of World Congress on Nature Biologically Inspired Computing (NaBIC) 2009*, 2009, pp. 1552–1555.

[152] S. Young *et al.*, *The HTK book (for htk version 3.2.1)*. Cambridge University Engineering Department, 2002.

List of candidate's work related to the thesis

Journals (Impact)

- No publications

Journals (Reviewed)

- J. Bartošek, "A pitch detection algorithm for continuous speech signals using Viterbi traceback with temporal forgetting", English, *Acta Polytechnica*, vol. 51, no. 5, pp. 8–13, 2011, ISSN: 1210-2709

ISI Web of Knowledge (WoS) excerpted publications

- J. Bartošek and V. Hanžl, "Foot detection in Czech using pitch information and HMM", in *Proc. of Text, Speech, and Dialogue 2013*, ser. Lecture Notes in Computer Science, Berlin: Springer, 2013, pp. 272–279, ISBN: 978-3-642-40584-6

  The paper has been cited in:

  - S. Sarma and U. Sharma, "Word boundary detection in Assamese language", *International Journal of Computer Sciences and Engineering (IJCSE)*, vol. 3, pp. 150–156, 1 Feb 2015, ISSN: 2347-2693

- J. Bartošek and V. Hanžl, "Intonation based sentence modality classifier for Czech using artificial neural network", English, in *Advances in Nonlinear Speech Processing*, ser. Lecture Notes in Computer Science, Heidelberg: Springer, 2011, pp. 162–169, ISBN: 978-3-642-25019-4

  The paper has been cited in:

  - A. Warsi, T. Basu, and D. Mazumdar, "Role of prosody in automatic modality recognition of bangla speech", in *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association*, Portland, Oregon, USA, 2012, pp. 675–678
  - J. Volín, L. Weingartová, and O. Niebuhr, "Between recognition and resignation – the prosodic forms and communicative functions of the Czech confirmation tag "jasně"", in *Proceedings of the 7th International Conference on Speech Prosody*, Dublin: TCD, 2014, pp. 115–119

- J. Bartošek and V. Hanžl, "Exploring abilities of merged normalized forward-backward correlation for speech pitch analysis", English, in *Proc. of International Conference on Applied Electronics*, Plzeň: Západočeská univerzita v Plzni, 2011, pp. 35–38, ISBN: 978-80-7043-987-6

Other publications

- J. Bartošek, "Pitch detection algorithm evaluation framework", English, in *Proc. of 20th Czech-German Workshop on Speech Processing*, Prague: Institute of Photonics and Electronics AS CR, 2010, pp. 118–123, ISBN: 978-80-86269-21-4

  The paper has been cited in:

  - F. Tamburini, "Una valutazione oggettiva dei metodi più diffusi per l'estrazione automatica della frequenza fondamentale", in *Atti dell IX Convegno Nazionale dell'Associazione Italiana di Scienze della Voce (AISV2013)*, Bulzoni:Roma, 2013, pp. 427–434, ISBN: 978-88-7870-901-0

- J. Bartošek, "Real-time možnosti zkrácené autokorelační funkce pro detekci základní frekvence", Czech, in *Proc. of 20th Annual Conference Proceeding's Technical Computing*, Bratislava, 2012, pp. 1–8, ISBN: 978-80-970519-4-5

- J. Bartošek, "Czech Text-to-Foot converter", in *Proc. of POSTER 2013 - 17th International Student Conference on Electrical Engineering*, Prague, 2013, ISBN: 978-80-01-05242-6

- J. Bartošek and V. Hanžl, "Comparing pitch detection algorithms for voice applications", English, in *Proc. of Digital Technologies 2010*, Žilina, SK, 2010, ISBN: 978-80-554-0304-5

- J. Bartošek and V. Hanžl, "Prozodie a modelování přízvukových taktů", Czech, in *III. Letní Doktorandskeé Dny 2013*, Original document in Czech, Praha, CZ, 2013, pp. 98–101, ISBN: 978-80-01-05251-8

- J. Bartošek and V. Hanžl, "Prozodická analýza a modelování", Czech, in *LETNÍ DOKTORANDSKÉ DNY 2012*, Praha, CZ, 2012, pp. 78–80, ISBN: 978-80-01-05050-7

- J. Bartošek, "Porovnávání algoritmů pro detekci základní frekvence se zaměřením na řečové signály", Czech, in *Analýza a zpracování řečových a biologických signálů - sborník prací 2010*, Praha, CZ, 2010, pp. 1–8, ISBN: 978-80-01-04680-7

- J. Bartošek, "The use of prosody in speech recognition systems, punctuation detector for czech speech", English, in *Králíky 2010*, Brno, CZ, 2010, pp. 24–27, ISBN: 978-80-214-4139-2

- J. Bartošek, "Prozodie, zjištění a využití základního tónu v rozpoznávání řeči", Czech, in *Analýza a zpracování řečových a biologických signálů - sborník prací 2009*, Praha, CZ, 2009, pp. 1–8, ISBN: 978-80-01-04474-2

Applied results or patents

- No applied results or patents were created

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

- J. Rusz, R. Čmejla, J. Bartošek, *et al.*, "Assessment of voice and speech impairment", in *Workshop 2011,CTU Student Grant Competition in 2010 (SGS 2010)*, Praha: ČVTVS, 2011, pp. 1–6