Foot Detection in Czech Using Pitch Information and HMM

Jan Bartošek and Václav Hanžl

Department of Circuit Theory, FEE CTU in Prague,
Technická 2, 166 27 Praha 6 - Dejvice, Czech Republic
{bartoj11,hanzl}@fel.cvut.cz
http://obvody.fel.cvut.cz

Abstract. In the presented work we are dealing with modelling and detection of lexical stress-group (foot) for Czech language. Detection of foot as one type of supra-segmental (prosody) information nearly corresponds to detection of word boundaries. Every native speaker is able to distinguish the feet in continuous speech, but on the other hand there are still no obvious connections between the sound qualities (pitch, intensity, syllable length) and foot prominence realization in Czech. In the experiment we tried to train the Hidden Markov Models (HMM) for Czech feet representation using only pitch information in the syllable nuclei. The most of Czech SPEECON database was used as an experiment source database. A necessary part of the presented system is a tool that transforms given Czech text into the foot units according to the known linguistic rules.

Keywords: prosody, stressed-group detection, foot, pitch, clitics absorption, ASR, HMM.

1 Introduction

Prosodic information is still not sufficiently used in nowadays automatic speech recognition (ASR) systems. Although this kind of system can generally benefit from the use of prosody, it is commonly lost during the parametrization process. Besides modality detection of the sentence, the lost prosodic information should be also able to give a cue about borders of stress-group units (feet). This could in the end help the ASR to decide in special cases where syllable chain of the utterance is ambiguous without foot placement decision (typical Czech examples are “proti vnějším“ vs. “protivnějším“ or “světlo v ní mají“ vs. “světlo vnímají“).

There have been several attempts to find stressed syllables in the utterance. Probably one of the first was study [1] which was dealing with relevant factors as prominence indicators. For fixed stressed language there is study [2] developing word boundary detection system for Hungarian with trained HMM on F0 and energy as prosodic features obtained in regular time interval. In Hungarian, all acoustic qualities realizing lexical stress correspond and presented success rate was about 77% for word boundary detection task. Study [3] engaged in detection of emphasized words for Czech, but in that case it was sentential stress detection rather than lexical stress we are dealing with. They claim that stressed syllables in Czech are generally characterized mostly by intensity

I. Habernal and V. Matousek (Eds.): TSD 2013, LNAI 8082, pp. 272–279, 2013.

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 out of 180 sentences recorded specially for this experiment. The most significant feature alone was found to be relative word prolongation (86% score). In [4] there was introduced a special system of stressed/unstressed vowels for Dutch acoustic modelling in their ASR, but the final decrease of WER was not observed. Work [5] followed up the stress detection using the spectral slope as the feature. They recorded three-syllable pseudo words with different option of stress placement. By computing the energy ratios of bands 350-1100Hz and 2300-5500Hz they observed that all stressed vowels show less spectral scope value, they are more spectral-balanced than the unstressed. Unfortunately, their approach is dependent on a knowledge of particular vowels.

According to known research [6,7], the Czech prosodic system for lexical stress realization is very unique and mentioned approaches [2] are not directly applicable for Czech. We do not know about any research that explores detection success rate of Czech lexical stress (or foot units) on larger data corpus based on various acoustic qualities. In contrast to all mentioned works for Czech, we operate on huge set of utterances from Czech SPEECON database used commonly for training of ASR systems for Czech. In this initial experiment we focus on the possibility of stress-group (foot) unit detection based purely on melody (pitch) information. We also focus on development of an automatic syntactic-level ASR module with ability of utterance division into stressed-groups. Our Czech feet modelling approach is based on collaboration with the ASR that provides specific information (especially syllable nuclei centers time-stamps obtained by force-alignment process) in combination with utilizing the raw acoustic data for the feature extraction. In this paper we tried to find how pitch information as the only feature used can model feet in Czech.

The article is organized as follows: Section 2 brings some theory about Czech feet. Used data set is described in Section 3, used features and their normalizations in Section 4. An overview of the whole training phase of the system for Czech feet detection is presented in Section 5. Experimental setup is closely presented in Section 6 followed by achieved results (Section 7), which are discussed together with possible future work.

2 Facts about Czech Language and Feet Realizations

A foot (stressed group) is a unit of speech binding one prominent syllable and certain count of non-prominent syllables. Czech is a fixed-stress language with stressed first syllable. From the speech rhythm point of view Czech is ”syllable-timed” language (syllable is considered as basic perceived time resolution step of the speech), which means there is usually no speeding up or slowing down of the speech according to the length of the foot. In [6] it is claimed that foot is really perceived as a basic unit of the utterance. Prominence in Czech is probably achieved by mixture of sound qualities - melody contour, speech dynamic and also durational features of syllables (especially their vowels), but there still has not been discovered definitive description of how these

sound quality features correspond together when creating the prominence (in contrast with [2] where speech dynamic and also pitch have similar contours with peaks corresponding to prominent syllables). Description of each type of prominence realization follows:

a) When prominence is realized by pitch change, it can be done by both rising or lowering the pitch in Czech. There are known typical musical intervals for both cases that represent valid pitch prominence (measured on synthetic isolated word database): up - by one semitone, down - by four semitones (musical third). When there is longer chain of syllables in the foot, the pitch changes are even less (level of musical quartertones). This fact involves the need of a precious pitch detection algorithm (PDA) when studying foot intonation contours. Pitch prominence is probably used most often as word-level prominence indicator, but greater pitch changes than denoted are probably perceived as key sentence melody events. This involves the problem - sentence modality is distinguished by the melody contour and this occurs also across the foot prominence realized also by the pitch. Resulting pitch contour of the sentence is thus a combination of foot pitch prominences and sentence melody and when examining the pitch contour, we are dealing with both sentence and word-level segmentation. One of the latest studies [7] about sound qualities of the feet prominence and the role of the pitch in determining foot boundaries for Czech confirms one of the former theories. The strongest tendency found is that the pitch contour with clear F0 drop in the middle does not represent an acceptable form of a foot.

b) In the past it was thought that the prominence in Czech is done only by the dynamic. This has been proven to be true for isolated words synthetic material, but according to the later studies aimed at real complex utterances it very often occurs, that the syllable with prominence is less dynamic than the rest of the foot. That is why the dynamic is apparently not the driving feature of the prominence.

c) Length of syllables: In Czech it is not allowed to evidently prolong or shorten the syllable duration, because of its functional meaning. Nevertheless, very slight changes in the duration might also play a role in determining the prominence.

In continuous speech Czech multi-syllable words very often keep their independence and they create own feet. On the other hand single syllabic words very often lose their autonomy and create one foot together with the other neighbouring word. These single syllabic words are called clitics. There exist grammatical rules [8] how particular word categories behave - if they join their predecessor (then they are marked as enclitics) or successor (proclitics). On the basis of these rules was created lexical module capable of text-to-foot conversion [9].

3 Used Data Set

Firstly, we considered to use the Czech Audiobooks as the corpus for the whole system but we decided not to do this. The main reason was the professionalism of the speakers which lead into excessive intonation over the whole data set. This fact can be beneficial for sentence modality detection, but for foot detection this can be inconvenient. Rather we decided to use part of the standard Czech SPEECON speech database with recordings of 550 adult non-professional speakers (16kHz, 16-bit, mono). We used cleaned

subset of sentences and further filtered diftongs because of their ambiguous syllabification in Czech. We finally obtained final subset of 10,022 sentences which were then processed to obtain their foot division and features. Sentences were then divided into training and testing subset (ratio 9/1).

4 Used Features and Their Normalizations

In this experiment we limited on using this set of features computed at each syllable nucleus: pitch, Δpitch and $\Delta\Delta$pitch. The features are not extracted equidistantly on the time axis, but are driven by the occurrence of the syllable nucleus. To obtain pitch in musical units we firstly converted fundamental frequencies from Hz into semitones scale related to the mean pitch frequency for the utterance measured across voiced regions only (Eq.1). We further refer to this type of normalization as "norm0".

$$SemitoneDifference = 12 \log_2(\frac{F0}{F0_{mean}}) \quad (1)$$

Second 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. Third 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) tends to make the curve more convex for the other parts of the utterance. This is why we introduced "norm3" where we do the same process as for "norm2" but with removed last foot of the utterance from all data structures.

5 System Overview

The process of training (and also testing) data preparation can be seen in the Fig.1. Czech SPEECON sentence subset is the only information source for further processing and feature extraction. Firstly, fundamental frequency (F0) of all audio files are obtained using Praat [10] autocorrelation function with lowered voice/unvoiced threshold (set to value 0.2). This threshold is quite low for ordinary use, but is more convenient for our case as we know exactly where the syllable nuclei should be located (and thus we allow some VUV errors outside our regions of interest). Praat F0 output is converted into the form with 1ms timestep for convenient further processing.

Another step is a retrieval of the .NUC files with timestamps of syllable nuclei. Using HTK we created and trained standard context dependent triphone acoustic models (three states per model) on the same dataset that were further used for force-alignment of all the utterances. This allowed us to obtain syllable nuclei time-stamps as the aligned times of triphone middle states. Stressed and non-stressed vowels were not anyhow distinguished in the phoneme alphabet used for force-alignment.

Having both .F0 and .NUC files we can create feature vector (.TRAIN) files for model training. A production of these feature files is not limited to just picking the

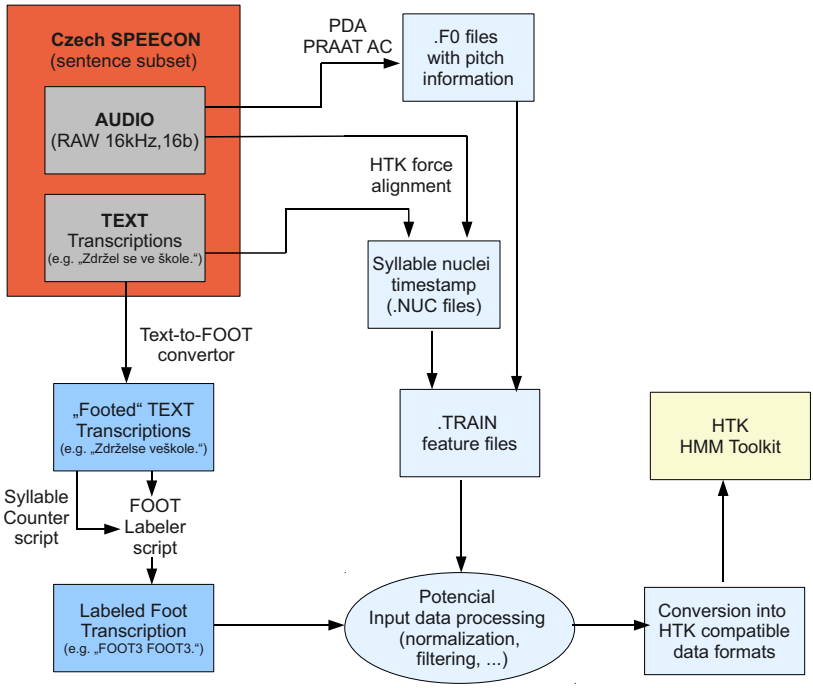


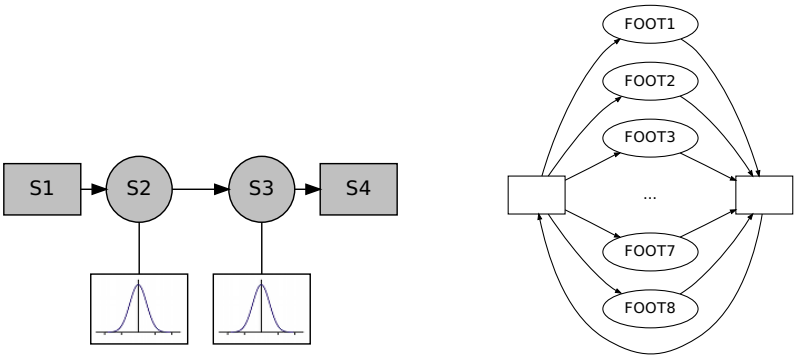
Fig. 1. Scheme of the training data preparation process

corresponding pitch in the time of nucleus center occurrence. It does also contain a logic for repairing the estimates of these centers by seeking for the contiguous regions of pitch around nucleus center (influencing the range for delta and double delta features computation) and also logic for approximation of missing pitch (frame is denoted as unvoiced by PDA). Moreover, various types of pitch normalization can occur during and after creation of .TRAIN files.

On the lexical side of the chain utterance text transcriptions enter the Text-to-Foot (TTF) module [9]. Basically, it follows the strong rules for Czech clitics absorption [8]. The transcription is converted by the module into the output text containing stressed groups division (an example of Czech sentence on the input and its foot-oriented conversion can be seen in the Fig.1). It has been proven [9], that used TTF module generates (after transformation of the whole dataset) very similar distribution of feet with given lengths as it is referenced by respected literature [6] for Czech speech. Obtained stress-grouped transcription enters the FOOT labeller which in connection with the syllable counter module produce final FOOT-labeled utterance transcription. One of possible outputs can be e.g, "FOOT2 FOOT3." (one 2-syllable foot followed by one 3-syllable foot). The syllable counter can manage special diftongs as well as syllable-creating consonants 'l' and 'r'.

6 Experiment Setup

To model our problem we utilized Hidden Markov Model (HMM) approach, because our task is similar to standard speech recognition tasks, where HMM framework is widely adopted. All the experiments were performed using HTK Speech Recognition Toolkit [11]. We trained HMMs for various feet lengths (1-8 syllables) on the training subset. Our models of feet are in comparison with standard speech triphone models without state self-loops and backward state transitions. Thus, we do not allow the model to stay in any state and state-flow of our system is strictly forward with coming input features. We illustrate HMM model for two-syllable stressed-group (label FOOT2) in the Fig.2(a). Each emitting HMM state was modelled using one mean and variance (no mixtures). Each utterance is generally modelled by the grammar depicted in the Fig.2(b). The a priori probability was not modified for any of the models and thus is equal for all of them.



(a) HMM prototype of two-syllable FOOT2 (b) Used Grammar

Fig. 2. Illustration of used HMM modelling

We trained and tested the models using various forms of pitch data, but main experiments were related to four suggested types of normalization. As for testing with ”norm1“ data, we realized that in real system these data will not be available, because the division of the utterance into feet will be unknown. This is why we decided to make experiments with feet models trained on ”norm1” data, but tested with different normalization types available in real situation. We also prepared version of experiment with filtered final feet of utterances. We expected this could improve the accuracy of the system, because we would remove the feet most influenced by sentence intonation. Besides, in another version we filtered out all the sentences that contained comma indirect speech, because their feet should be mostly affected by complex sentence intonation.

7 Results, Discussion and Future Work

We performed various versions of experiment, but only those most valuable are quoted in Tab.1. We actually found that neither filtering of the last foot nor filtering the complex-sentences out of the dataset did not improve the accuracy. By using “norm1“ feature

Table 1. Results from HResults for foot detection using pitch information

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

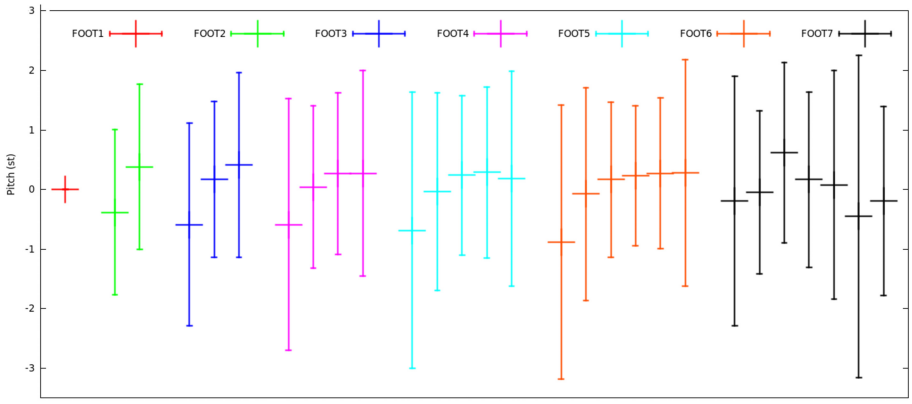


Fig. 3. Pitch means and STD for "norm1" trained HMM models of Czech feet

files, we were able to obtain accuracy up to 46%, which denotes rightness of this type of normalization for our task. In the Fig.3 we can see these "norm1" trained models FOOT1-FOOT7 with plotted pitch values of emitting states. Models for FOOT2-FOOT6 very well satisfy the theoretical condition for foot existence declared in [7] – that intonational contour with pitch drop in the middle does not create acceptable form of foot. In more real scenario using norm3 features as testing data while still keeping HMM models trained on norm1 features, we were able to recognize feet in utterances with 32% accuracy.

Although reached results are not much impressive, there is still a lot of possible future work and improvements pending:

1. Collaboration with Czech language specialists (phonetics), which can lead into careful verification of used features and corresponding labels and also improvements of the used text-to-foot module
2. Choice of features - except the pitch there are another features that are obtainable and worth to try (intensity, durations of syllables or vowels and spectral slope [5]).

8 Conclusions

We have created the framework for the experiments dealing with the automatic Czech feet detection based on the subset of the Czech SPEECON database. On the lexical level we utilize necessary module that converts given sentence into feet units. In this experiment, stressed-groups HMM models were trained using only pitch information, but with various types of normalizations. Trained models are in strong accordance with results of other phonetic studies examining Czech feet pitch behaviour [7]. They statistically confirm the fact that most of lexical stress is realized by pitch drop on the first syllable of the foot. Unfortunately, on our testing set we were able to achieve only 32% foot recognition accuracy. To achieve better scores, it seems that additional features extracted from the acoustic signal would be needed.

Acknowledgments. This work has been supported by the GA of the Czech Technical University in Prague, grant No. SGS12/143/OHK3/2T/13.

References

1. Fry, D.B.: Experiments in the perception of stress. Language and Speech 1, 126–152 (1958)
2. Vicsi, K., Szaszák, G.: Using prosody to improve automatic speech recognition. Speech Commun. 52, 413–426 (2010)
3. Kroul, M.: Automatic detection of emphasized words for performance enhancement of a czech asr system. In: Proceedings of 13th International Conference Speech and Computer (Specom 2009), Petersburg, Russia, pp. 470–473 (2009)
4. Kuijk, D.V., van den Heuvel, H., Boves, L.: Using lexical stress in continuous speech recognition for dutch. In: Proc. ICSLP 1996, pp. 1736–1739 (1996)
5. Volín, J., Zimmermann, J.: Spectral slope parameters and detection of word stress. In: Proceedings of Technical Computing Prague (Humusoft), Praha, pp. 125–129 (2011)
6. Palková, Z.: Fonetika a fonologie češtiny (Phonetics and phonology of Czech). Karolinum, Praha (1994)
7. Palková, Z., Volín, J.: The role of f0 contours in determining foot boundaries in czech. In: Proceedings of the 15th ICPhS, Barcelona, vol. 2, pp. 1783–1786 (2003)
8. Hauser, P.: Základy skladby češtiny. Masarykova Univerzita, Brno (2003)
9. Bartošek, J.: Czech text-to-foot converter. In: POSTER 2013 (CD-ROM) (2013)
10. Boersma, P.: Praat, a system for doing phonetics by computer. Glot International 5, 341–345 (2001)
11. Young, S.: The htk hidden markov model toolkit: Design and philosophy. Entropic Cambridge Research Laboratory, Ltd. 2, 2–44 (1994)