The Czech Speech and Prosody Database Both for ASR and TTS Purposes

Jáchym Kolář, Jan Romportl, Josef Psutka

Department of Cybernetics, University of West Bohemia, Univerzitní 8,
306 14 Plzeň, Czech Republic
jachym@kky.zcu.cz, rompi@students.zcu.cz, psutka@kky.zcu.cz

Abstract

This paper describes a preparation of the first large Czech prosodic database which should be useful both in automatic speech recognition (ASR) and text-to-speech (TTS) synthesis. In the area of ASR we intend to use it for an automatic punctuation annotation, in the area of TTS for building a prosodic module for the Czech high-quality synthesis. The database is based on the Czech Radio&TV Broadcast News Corpus (UWB_B02) recorded at the University of West Bohemia. The configuration of the database includes recorded speech, raw and stylized F_0 values, frame level energy values, a word- and phoneme-level time alignment, and a linguistically motivated description of the prosodic data. A technique of prosodic data acquisition and stylization is described. A new tagset for a linguistical annotation of the Czech prosody is proposed and used.

1. Introduction

In this paper we describe our experience with the preparation of a large Czech prosody database. Prosody and prosodic features may be useful in a broad spectrum of branches related to computer speech processing. The most natural area of use is, of course, the TTS synthesis. But also in other areas, related rather to the ASR, the prosody could be a very helpful cue. One important area is spoken dialog systems, where the prosody can be used for dialog act recognition as well as in parsing or a fast end-of-utterance detection. A second area is the annotation of a structure in speech, including punctuation, disfluencies and topic segmentation. The database described in this paper should be primarily used in an automatic punctuation annotation and in a building of a prosodic module for the Czech high-quality TTS synthesis. The prosody database is based on the Czech TV&Radio Broadcast News Corpus.

In automatic punctuation, the goal is to improve readability of an automatic transcription, and/or to arrange it into a form fitting more in natural language processing and information retrieval. As the ASR systems made a big step ahead in recent years, a large volume of audio data can now be transcribed automatically. However, it is very difficult for humans to read these transcripts because of missing sentence boundaries, punctuation marks and casing. In recent years, several approaches to automatic punctuation exploiting lexical, acoustic, and prosodic features have been examined [1, 2, 3].

For TTS synthesis purposes, we are interested in designing an estimator of prosodic characteristics which would be able to assign appropriate (e.g. as much naturally sounding as possible) melody, intensity and timing to an arbitrary input text. To estimate the prosodic characteristics we need adequate data.

Station	M	F	L	N	G	PN	PC	S
Nova	TV	30	25	2	M,F	Y	Y	Y
Prima	TV	23	12	1	F	Y	Y	Y
ČT 1	TV	22	30	2	M,F	Y	Y	Y
ČRo 1	R	81	20	1–2	M,F	Y	Y	Y
ČRo 1	R	57	5	1	M,F	N	Y	Y
Praha	R	90	5	1	M,F	Y	Y	Y
F 1	R	30	5	1	M,F	N	Y	N
Vltava	R	14	5	1–2	M,F	N	Y	N

Table 1: *Sources of Broadcast News in UWB_B02 (where M denotes the broadcast medium, F number of files, L length of the program in minutes, N number of moderators in the studio, G genders of moderators, and flags PN, PC and S indicate whether prearranged news, phone calls, and street reporters could appear in the program.)*

2. Czech TV&Radio Broadcast News Corpus

The UWB_B02 [4] is the Czech TV&Radio Broadcast News corpus spanning the period February 1, 2000 through April 22, 2000. During this time news broadcasts on 3 TV channels and 4 radio stations were recorded. The whole corpus contains over 60 hours of audio stored on 347 waveform files, which yield about 26 hours of pure transcribed speech. The broadcast sources and their form are described in Table 1.

The broadcast news does not contain the weather forecast, the sports news and the traffic announcements. The signal is single channel sampled at 44.10 kHz with 16-bit resolution, but for this database preparation purpose we have used waveforms downsampled to 22.05 kHz. Some interesting numbers related to the UWB_B02 corpus can be found in Table 2. Details about corpus annotation are given in [4].

#sentences	16,483
#turns	5,922
#tokens	233,959
#tokens per sentence	14.19
#distinct words	31,936
#speakers	284
#male speakers	188
#female speakers	96
#sentences by males	9,972
#sentences by females	6,511

Table 2: *Some numbers related to the UWB_B02 corpus*

3. Prosody database

The prosody database was created from the corpus stated above. We had audio files, their transcriptions and vocabulary at disposal, whereof we had to create the prosody database. We were interested in prosodic features such as pitch, intensity, speaking rate, and in linguistical description of prosodic data. So the prosodic database should contain:

- Speech signal divided into suitable units and its phonetic annotation.
- Raw and stylized F_0 contour, frame level energy values.
- Word and phoneme time alignments. Phoneme duration statistics.
- Linguistically motivated description of the data.

The database has to be designed in such a way so that we can compute values of any desired set of prosodic features.

3.1. Data segmentation and transcripts

At first we had to divide the large audio files into smaller units. With a respect to the planned use of the database, we have chosen a turn as this unit. It could also be a spurt, which is defined as a stretch of contiguous speech containing no pauses longer than some defined limit (e.g. 0.5 seconds) [5]. The spurt might be a more convenient unit for the automatic punctuation, but is not a convenient unit for the TTS. Since in most of broadcast news programs recorded in the used corpus more than one moderator is in the studio, and announcements are often interrupted by various events (e.g. phone calls, jingles etc.), turns are not too long.

All turns marked as “Nontrans” in the transcripts were discarded from further processing. These are segments of the audio file containing no spoken content or the spoken content accompanied by noise or recorded in a lower quality (e.g. reporters on the phone). The marking of “Nontrans” segments was done manually during the corpus annotation [4]. As start and end times for each turn are available from the transcripts, we could easily split audio files into plenty of small files containing just one turn. Also the original transcripts had to be arranged. By parsing the original XML documents, we gained the HTK-style [6] master label file (MLF) where commas, periods, and question marks were replaced by <COM>, <PER>, and <QM> tags.

3.2. Word and phoneme alignments

Since prosodic features are commonly utilized at the word level, we are in need of knowing where each word starts and ends. Because we need to know those word boundaries as accurate as possible, we used a forced alignment procedure to generate them from the data. We have to note that for some purposes, such as “robust automatic punctuation”, alignments generated by the ASR could be useful too.

We trained in HTK new models on UWB_B02 data, and these models were used for the forced alignment. The alignment was generated for each segment (turn) at once. We also obtained phonemes and pauses durations from these alignments. Vowels duration statistics are shown in Table 3. The duration of vowels is more interesting than duration of consonants, since vowels influence more significantly the overall speaking rate.

3.3. F_0 and energy extraction and stylization

The major part of prosodic features is related to a F_0 contour, so its reliable extraction and stylization are essential. A lot of various methods for the F_0 extraction were proposed. We have used RAPT (Robust Algorithm for Pitch Tracking) [7], which is included in the *Snack* sound toolkit [8]. The frame level RMS energy values were also computed by using the *Snack* toolkit.

VOW	N_OCC	A_DUR	VAR	STD
a	87,532	65.23	859.43	29.32
aa	31,569	106.74	1,154.30	33.98
aw	545	124.48	890.05	29.83
e	121,417	61.46	620.33	24.91
ee	17,626	90.90	1,329.15	36.46
ew	42	153.33	1,274.60	35.70
i	81,645	63.41	786.47	28.04
ii	49,049	75.84	1,297.32	36.02
o	98,837	60.66	522.25	22.85
oo	697	115.94	1,012.64	31.82
ow	7,750	105.50	1,379.85	37.15
u	31,350	68.78	976.61	31.25
uu	9,075	102.58	2,218.79	47.10

Table 3: *UWB_B02 vowels duration statistics (where N_OCC denotes a number of occurrences, A_DUR average duration in milliseconds, VAR variance, and STD standard deviation)*

As raw pitch values are giving no sense of relative pitch rises and falls for a particular speaker, some kind of normalization is necessary [5]. Another reason for the normalization is the presence of octave errors. Although RAPT pitch tracking method is fairly robust, it is, as other methods, also liable to halving and doubling errors. As it was shown in [9], clean pitch has a lognormal distribution. The estimated pitch, which has been exposed to halving and doubling, could be fitted to the lognormal tied mixture (LTM) model with 3 components. Components of the mixture correspond to distributions of halved, accurate, and doubled pitch values. The lognormal tied mixture model for the estimated pitch distribution can be written as

$$\begin{aligned}&\log(\hat{F_0}) \sim LTM(\mu,\sigma,\lambda_1,\lambda_2,\lambda_3) = \\ &\lambda_1 \cdot \mathcal{N}(\mu - \log(2), \sigma^2) + \lambda_2 \cdot \mathcal{N}(\mu, \sigma^2) \\ &+\lambda_3 \cdot \mathcal{N}(\mu + \log(2), \sigma^2)\end{aligned} \tag{1}$$

where $\sum_{i=1}^{3} \lambda_i = 1$. The parameters of the model can be estimated by using the Expectation-Maximization (EM) algorithm. After estimating parameters of the LTM model, we are able to determine the baseline value for a particular speaker. It indicates the point at which the probability of an accurate pitch is equal to the probability of its halving. The pitch values falling below the baseline were excluded from the later processing. Also the values being probably doubled were excluded. Remaining accurate values were then filtered by the median filter with a neighborhood of size 5. The estimated baseline value can be used for the pitch-related prosodic features normalization. In the future work we will pay attention to find out whether this normalization is better than the normalization to the mean value.

We also have to deal with the phenomenon of microintonation. The tracked, median filtered, and halved/doubled values removed pitch contour still contains a lot of local fluctuations [10]. These fluctuations are involuntary on the speaker's part and related to the physiology of speech. A common way to remove the microintonation is to stylize the pitch contour by a piecewise linear function. The line fits better interpret pitch movements intended by the speaker.

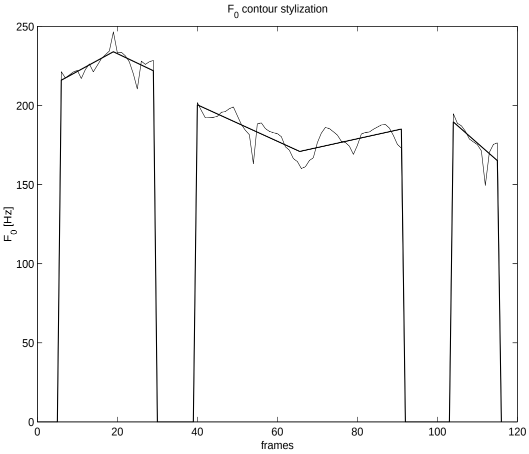

Figure 1: *An example of the F_0 contour stylization*

The number of lines per a particular voiced region is determined proportionately to its duration. The free parameters are coordinates of the piecewise linear function nodes $(x_k, y_k)_{k=0}^{K}$, excluding x-coordinates at the edges, x_0 and x_k, which are fixed. The fitting function is given by

$$g(x) = \sum_{k=1}^{K} (a_k x + b_k) \boldsymbol{I}_{\left[x_{k-1} < x \leq x_k\right]} \qquad (2)$$

where a_k is the slope and b_k is the intercept of the line defined by (x_k, y_k). The node parameters are estimated by optimizing the mean square error (MSE) between the pitch estimates and the stylized fit

$$MSE\left((x_k, y_k)_{k=0}^{K}\right) = \frac{1}{T} \sum_{t=1}^{T} (F_0(t) - g(t))^2 \qquad (3)$$

Nelder-Mead simplex method [11] was used for minimizing this objective function. An example of F_0 contour stylization by the piecewise linear model is shown in Figure 1.

Then it was necessary to line up the word boundaries with stylized F_0 contour nodes. Nodes that fell within a word's start and end are stored in an extensive database file, beside the word and phonemes duration values. Raw values of F_0 are also available. Therefore, we have the word-level prosodic features at hand.

3.4. Linguistical annotation

In addition to the above mentioned acoustically motivated prosodic data description the corpus is also enriched by the linguistically motivated description of the prosodic data (following the linguistical structuralism). Both of them are essential for our prosody module for the TTS system. Only turns spoken by a speaker who uttered in the corpus more than 100 sentences were annotated. This linguistical annotation consists of the following phases: sentence modality (a speaker's attitude towards a sentence in a communication, segmentation into phonemic words, segmentation into phonemic clauses, semantic accent, and prosodemes (abstract functional units). We propose a way and a new tagset for a linguistical annotation of the Czech prosody.

All phases are annotated manually using software tools to inspect the acoustic properties of given utterances. However, the most important tool is the annotator's hearing and since this level of annotation lies in the abstract sphere of the language description (with strong emphasis on one's perception), the annotators use the acoustically motivated data (such as the contour of F_0) only as a secondary tool, for example to help decide controversial cases. The annotated data are stored in the XML format.

3.4.1. Sentence modality

Each sentence contains a tag determining its modality. The corpus distinguishes the following sentence modalities (in the SENTENCE_MODALITY tag):

- declarative ("My name is John.")
- interrogative inquiring ("Are you going home?")
- interrogative supplementary ("When will they come?")
- imperative ("Do it now!")
- desiderative ("I wish it could happen!")
- exclamative ("You are joking!")

A specific tag indicates whether a sentence is or is not parenthetic.

3.4.2. Phonemic words

A phonemic word is a group of words subordinated to one word accent (stress). We have an algorithm for phonemic words detection in written text; obviously this algorithm does not take the acoustic characteristics of realized utterances and always proposes one of more possible variants of phonemic word placement (usually the most emotionally neutral). The text from the corpus is automatically pre-tagged by this algorithm, yet this often does not correspond to the phonemic words realized by speakers in the real utterances, hence the annotators manually correct this placement. The phonemic words are indiscerptibly bound to word accents (stresses) and so they are marked together by apostrophes preceding stressed syllables (the stress in the Czech language is fixed to the first syllable of a phonemic word).

3.4.3. Phonemic clauses

A phonemic clause is such a segment of speech where a certain intonation scheme is continuously realized. A single sentence often contains more phonemic clauses. The annotators place the borders of phonemic clause on the perceptional base. The tag C1 marks a phonemic clause and C2 marks a phonemic clause followed by a pause.

3.4.4. Semantic accent

By this term we call such a word (or sometimes even more words) in a sentence, which is emphasized (using acoustic means) by a speaker. It is marked by the SA tag preceding the word. We are actually not interested in "where a semantic accent is", but rather in "where a communionist (substituted by an annotator) thinks it is". If (according to an annotator) there is no emphasis in the sentence, the semantic accent is automatically placed on the last phonemic word. Moreover, it is very important to have non-isolated sentences, e.g. with some preceding and following context.

3.4.5. *Prosodemes*

Prosodeme is an abstract intonational pattern established in a certain function within the language system. We have postulated that any single phonemic clause consists of two prosodemes: so called “null prosodeme” and “functionally involved prosodeme” (the latter always starts at the semantic accent), depending on the communication function the speaker intends the sentence to have. We distinguish the following prosodemes:

P0 – null prosodeme
P1 – prosodeme terminating satisfactorily
  P1-1 – no indication
  P1-2 – indicating emphasis
  P1-3 – indicating imperative
  P1-4 – indicating interjection
  P1-5 – indicating wish
  P1-6 – specific
P2 – prosodeme terminating unsatisfactorily
  P2-1 – no indication
  P2-2 – indicating emphasis
  P2-3 – indicating “wh-” question
  P2-4 – indicating emphasized “wh-” question
  P2-5 – specific
P3 – prosodeme nonterminating
  P3-1 – no indication
  P3-2 – indicating emphasis
  P3-3 – specific

3.4.6. *The example*

(the tags for null prosodemes and semantic accents in their automated final position are omitted because of their redundancy)

<S ID=00001 MODALITY=DECLARATIVE >
<C1>’V posledních ’dnech ’naší <P3-1> ’dovolené </P3-1>
</C1><C2> jsem <P3-2><SA> ’konečně’pochopil
</P3-2> </C2><C2> že ’nesnáším <P1-1>’cestování
</P1-1></C2>//</S>.
(lit.: In the last days of our holiday I finally realized that I hate traveling.)

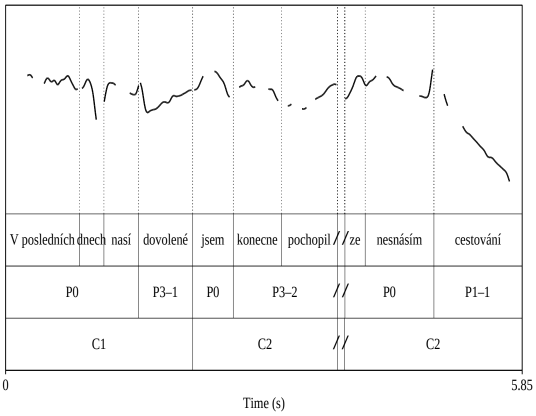

Figure 2: *An illustration of the linguistical annotation*

4. Conclusion

We have prepared a large Czech prosody database based on the Czech TV&Radio Broadcast News Corpus. The database contains both acoustically and linguistically motivated description of the prosodic data. We have proposed and used a new tagset for a linguistical annotation of the Czech prosody. The database has a form which allows us to compute values of any desired set of prosodic features.

We plan to use this database in an automatic punctuation annotation, and also in designing an estimator of prosodic characteristics for the high-quality Czech TTS system. In the future we also intend to develop a similar database from a spontaneous speech corpus.

5. Acknowledgement

Support for this work was provided by the Ministry of Education of the Czech Republic, project No. MSM 235200004.

6. References

[1] Ji-Hwan Kim, Woodland, P.C.: The Use of Prosody in a Combined System for Punctuation Generation and Speech Recognition. Proceedings of EUROSPEECH 2001, p. 2757–2760, Aalborg Denmark, 2001

[2] Jing Huang, Zweig, G.: Maximum Entropy Model for Punctuation Annotation from Speech. Proceedings of ICSLP, p. 917–920, Denver, 2002

[3] Baron, D., Shriberg, E., Stolcke, A.: Automatic Punctuation and Disfluency Detection in Multi-party Meetings Using Prosodic and Lexical Cues. Proceedings of ICSLP, p. 949–952, Denver, 2002

[4] Psutka, J., Radová, V., Müller, L., Matoušek, J., Ircing, P., Graff, D.: Large Broadcast News and Read Speech Corpora of Spoken Czech. Proceedings of EUROSPEECH 2001, p. 2067–2070, 2001

[5] Baron, D.: Prosody-Based Automatic Detection of Punctuation and Interruption Events in the ISCI Meeting Recorder Corpus. MSc. Research Project, University of California, Berkeley, 2002

[6] Young, S. et al.: The HTK Book (for HTK Version 3.1). Cambridge University, 2002

[7] Talkin, D.: A Robust Algorithm for Pitch Tracking (RAPT). In Speech Coding and Synthesis, Elsevier Science, Amsterdam, p. 495–518, 1995

[8] Sjölander, K.: The Snack Sound Toolkit. Available at http://www.speech.kth.se/snack/

[9] Sönmez, K., Heck, L., Weintraub, M., Shriberg, E.: A Lognormal Tied Mixture Model of Pitch for Prosody-Based Speaker Recognition. Proceedings of EUROSPEECH 97, p. 1391–1394, Rhodes, 1997

[10] Sönmez, K., Shriberg, E., Heck, L., Weintraub, M.: Modeling Dynamic Prosodic Variation for Speaker Verification. Proceedings of ICSLP, p. 3189–3192, Sydney, 1998

[11] Lagarias, J.C., Reeds, J.A., Wright, M.H., Wright, P.E.: Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions. SIAM Journal of Optimization, 9(1): p. 112–147, 1998