

Impact of Irregular Pronunciation on Phonetic Segmentation of Nijmegen Corpus of Casual Czech

Petr Mizera¹, Petr Pollak¹, Alice Kolman² and Mirjam Ernestus³

¹ Faculty of Electrical Engineering, Czech Technical University in Prague
{mizerpet,pollak}@fel.cvut.cz

² Radboud University Nijmegen & Christian University of Applied Sciences CHE
akolman@che.nl

³ Radboud University Nijmegen & Max Planck Institute for Psycholinguistics
mirjam.ernestus@mpi.nl

Abstract. This paper describes the first pilot study of phonetic segmentation applied to Nijmegen Corpus of Casual Czech (NCCCz). This corpus contains informal speech of strong spontaneous nature which influences the character of produced speech at various levels. This work is the part of wider research related to the analysis of pronunciation reduction in such informal speech. We present the analysis of the accuracy of phonetic segmentation when canonical or reduced pronunciation is used. The achieved accuracy of realized phonetic segmentation provides information about general accuracy of proper acoustic modelling which is supposed to be applied in spontaneous speech recognition. As a byproduct of presented spontaneous speech segmentation, this paper also describes the created lexicon with canonical pronunciations of words in NCCCz, a tool supporting pronunciation check of lexicon items, and finally also a minidatabase of selected utterances from NCCCz manually labelled on phonetic level suitable for evaluation purposes.

Keywords: spontaneous speech; casual speech; pronunciation reduction; phonetic segmentation; NCCCz

1 Introduction

In the past decades, speech technology applications have started being focused on the processing of spontaneous and informal speech which can be seen e.g. in automated transcription of various informal recordings from meetings or transcription of TV or broadcast programs for on-line subtitling or for archiving purposes. Due to this fact, researchers have become increasingly interested in the characteristics of spontaneous and casual speech in the most important world languages such as German, Dutch or English [1], [2], [3] and first steps have been taken in this field also for Czech.

The current speech recognition systems usually work very precisely for standard speech and we can find many works describing such systems for all world languages including Czech [4], [5]. However, the accuracy of speech recognition of

strongly spontaneous speech is significantly lower and the amount of published works describing the analysis or recognition of informal speech is also smaller. In this paper we present the first pilot study of phonetic segmentation accuracy applied to Nijmegen Corpus of Casual Czech (NCCCz), which was created to bring a missing corpus of Czech containing high-quality recordings from naturally occurring interaction which are suitable for detailed analysis of spontaneous speech in Czech [6].

The paper is organized as follows: firstly, the brief description of NCCCz is presented, secondly, the creation procedure of the lexicon with regular canonical pronunciations for NCCCz data is mentioned (together with the description of the tool supporting this step), and finally the results of the first analyses of the phonetic segmentation accuracy applied to casual Czech speech are presented.

2 The Nijmegen Corpus of Casual Czech

For the development of standard recognition systems, corpora of read speech or speech produced during formal interviews are used, e.g. SPEECON and Speech-Dat database [7], [8], [9]. The Nijmegen Corpus of Casual Czech (NCCCz) used in this work contains more than 30 hours of high-quality recordings of casual conversations among 10 triplets of male and 10 triplets of female friends. Always one speaker from each triplet acted as a confederate who asked two friends of the same gender (henceforth the naive speakers) to participate in recordings of natural conversations. The recording procedure was controlled by an experimenter.

Each session was recorded in a soundproof booth and in the first part of the recording, the confederate pretended to have received an important phone call that had to be answered immediately and the two naive speakers were left alone without information about whether they were already being recorded. Depending on the liveliness of the conversations between the two naive speakers, the confederate returned to the booth. Then the second part of the recording started, which consisted of free conversation among the three speakers. Various topics including school, relationships, common hobbies, and stories about all sorts of encounters were addressed. In the third part of the recordings, the experimenter entered the room with a list of questions on political and social issues and the speakers were asked to discuss at least four issues from the list and negotiate a common opinion for each question. The speakers were engaged in conversations approximately for 90 minutes and the recordings obtained by the above mentioned procedure contain very informal spontaneous data.

All speakers were recorded simultaneously on separate audio channels using cardioid microphones avoiding possible cross-talks in particular channels for each speaker. The whole corpus has been annotated at orthographic level using standard non-reduced transcription joined by additional marks for non-speech events. Corpus is freely available on demand as it is described in more details in [6].

3 The NCCCz lexicon

For the purpose of further studies and developments, the orthographic transcription of records had to be completed by the pronunciation lexicon. It always represents an important component which has significant impact on the accuracy of target ASR system. It is more important especially in the case of spontaneous or casual speech recognition where the process of coarticulation, assimilation and reduction often appears. Therefore the pronunciation lexicon should contain as many pronunciation variants as possible to capture this variability in informal spontaneous speech.

Due to very informal speaking style yielding to the presence of many rare and non-standard words, the lexicon of regular canonical pronunciation was created manually in cooperation with Czech native speakers with background in phonetics. The development of proper pronunciation lexicon can be described in the following steps: the first version was created using the rules of conversion from Czech orthographic transcription to canonical pronunciation [10]. The automatically generated pronunciation contained large amount of incorrect pronunciations mainly for foreign or the above mentioned non-standard words, but also due to poor voicing assimilation, phone softening, etc.

The correction (editing) of the pronunciation lexicon was supported by the extraction of information about the context of the given word form in the corpus and possibly also by listening to unclear words. For this purpose, we modified the *LexFix* tool for lexicon editing which provides the linking with both the orthographic transcription and audio signal to enable listening of recorded utterance.

The tool was created generally to support the determination of correct pronunciation of particular word forms. At the same time it was extended by other functions which simplify the work with a huge corpus. The possibility to search for the word in a huge corpus and display the neighbouring context can help determine the correct pronunciation. Typically, for foreign, rare, or generally non-standard words it may be difficult to decide about the pronunciation without the listening, so it is possible to play the particular sentence with the given word or play the whole displayed context. For more detail on how we worked with *LexFix* Fig. 1.

Eventually, the created pronunciation lexicon for NCCCz contains approx 30 000 word forms. During these checks, reduced pronunciations (e.g. “nějaký” vs. “ňáký”) were not marked so the current lexicon contains only canonical pronunciation with only a small amount of pronunciation variants. The reduction of the pronunciation is supposed to be solved automatically at further steps of this wider research, however, looking for the reduction rules is supposed to be supported also by this data from NCCCz together with the *LexFix* tool which enables the listening of particular occurrence of lexicon items in the corpus.

4 Phonetic segmentation

Automatic phonetic segmentation can be implemented in various ways. HMM-based automatic phonetic segmentation which is well-known as a forced align-



Fig. 1. Illustrative example of the work with the LexFix tool

ment is widely used technique. But other approaches for the phoneme localization using Bayesian changepoint detector, or artificial neural networks are also used by some authors [11], [12], [13].

The HMM-based forced alignment is a well known algorithm of looking for the maximum likelihood path through a composed acoustic model for the given utterance. Phone boundaries are then determined by the occupancy of HMM states representing particular phones over the found optimum path. The selection of proper pronunciations, ideally those which have been really realized in the given utterance plays a significant role in the segmentation accuracy. However, it is not often fulfilled in the case of casual speech.

The phonetic contents of each word used in the above mentioned algorithm is typically taken from the lexicon. Therefore, one important purpose of this work is to analyze the precision of phonetic segmentation in three basic cases, i.e. using three variants of pronunciation generation for the HMM-based forced alignment:

- using the lexicon with *canonical pronunciation*,
- using *actually realized reduced pronunciation* in each utterance which was transcribed manually,
- using the lexicon with *more pronunciation variants* containing several levels of pronunciation reduction.

This study should demonstrate by objective criteria (phonetic segmentation accuracy) the general impact of proper pronunciation selection in casual speech acoustic modelling which can then improve the accuracy of casual speech recognition (as it is shown illustratively in Fig. 2).

5 Experimental setup

HMM-based forced alignment was implemented by the open-source Kaldi Speech Recognition Toolkit [14] in a rather standard setup. As speech features we used

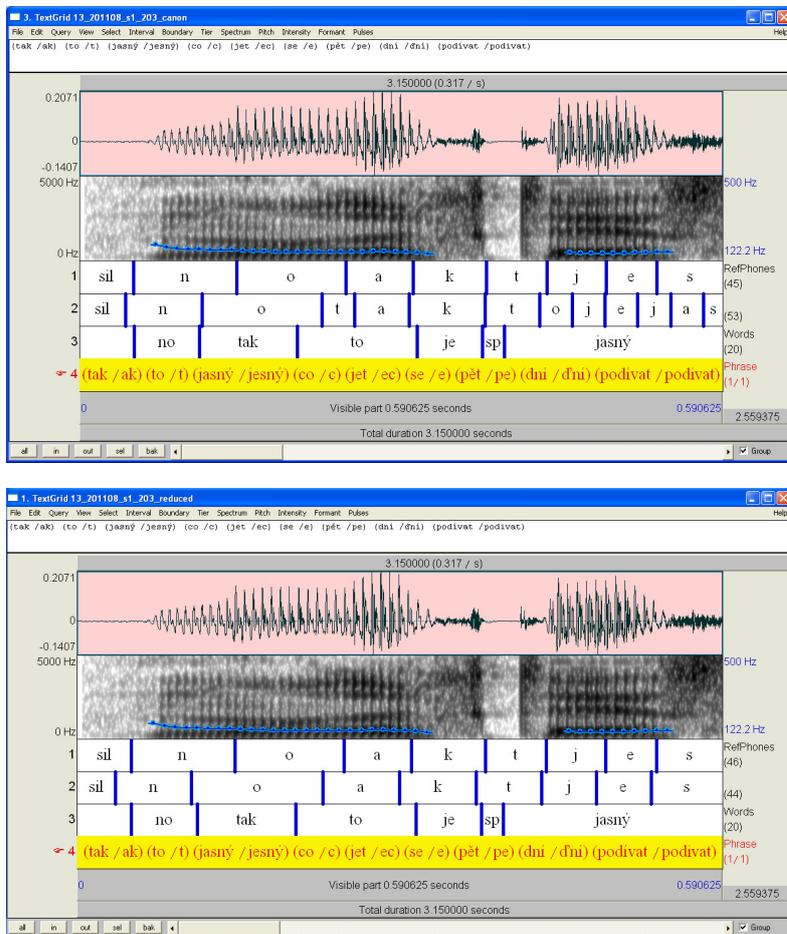


Fig. 2. Illustrative example of phonetic segmentation results with the canonical (top picture) and reduced (bottom picture) pronunciation.

commonly Mel-Frequency Cepstral Coefficients with the additional zeroth cepstral coefficient, completed by their delta and delta-delta features (MFFC_0_D_A). Cepstral mean normalization (CMN) was also applied to minimize the small mismatch between training and processed data. Short-time analysis used the frame length of 25 ms and frame shift of 10 ms. The GMM-HMM based acoustic model (AM) was based on triphones and trained on utterances from the Czech SPEECON database. The procedure of AM training was inspired by an example for the Wall Street Journal and the TIMIT database (KALDI recipes s4 and s5). The set used for the training of our AM contained utterances recorded in rather clean office environment and the amount was about 52 hours of read speech.

Finally, we used the *gmm-align* tool for forced alignment realization. As this tool produces the output of state-level alignments of utterances which are represented by transition-ids, we have created a simple tool for the conversion to

phoneme-level alignments of utterances. The output of this tool is in HTK MLF format, containing also the information on time boundaries of particular phones.

The experiments were carried out using selected utterances from the NCCCz corpus having the lengths of about 15 - 20 words. For this pilot experiment we have selected utterances of speakers with rather standard level of reduced pronunciation and we have selected data without further disturbance such as high background noise, high frequency of non-speech acoustic events or overlapping speech. This evaluation subset contains 19 utterances from 8 speakers which were now manually segmented at the phonetic level.

Eventually, in this evaluation subset we suppose to have at least 100 utterances from all speakers. The amount of data in this evaluation subset is summarized in Table 1 (values for the target state are estimated).

Sex	Current state				<i>Target state</i>			
	minutes	speakers	sentences	phones	<i>minutes</i>	<i>speakers</i>	<i>sentences</i>	<i>phones</i>
<i>Male</i>	1.09	6	16	923	<i>3.33</i>	<i>30</i>	<i>50</i>	<i>3000</i>
<i>Female</i>	0.20	2	3	172	<i>3.33</i>	<i>30</i>	<i>50</i>	<i>3000</i>
<i>Total</i>	1.29	8	19	1095	<i>6.66</i>	<i>60</i>	<i>100</i>	<i>6000</i>

Table 1. The NCCCz evaluation subset statistics

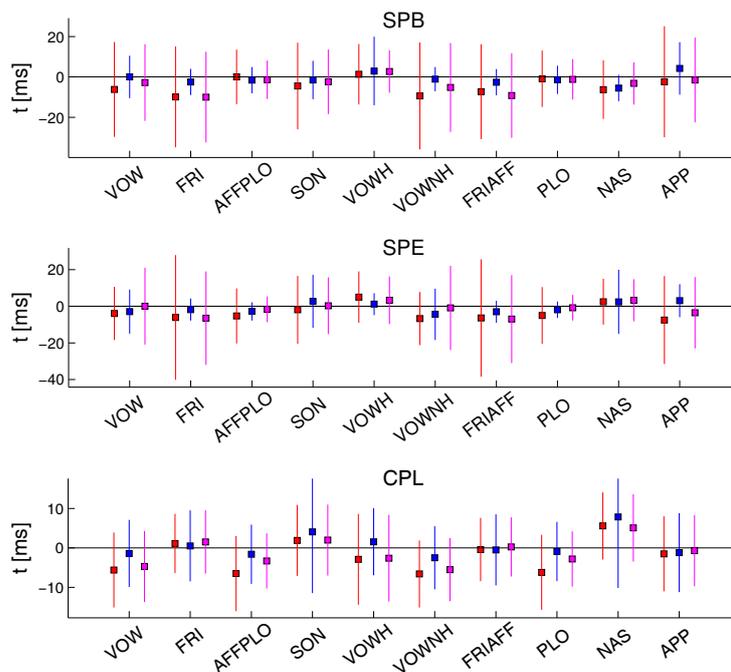
6 Results and discussion

This section presents the results of above discussed phonetic segmentation of casual Czech speech evaluation which were performed with various level of pronunciation reduction. The accuracy was quantified using the following criteria: *Shift of the Phone Beginning (SPB)*, *Shift of the Phone End (SPE)*, *Change of the Phone Length (CPL)*, and *Phone Error Rate (PER)*, more details are in [12].

The overall results are presented in the Table 2. The mean values and standard deviation of *SPB*, *SPE*, *CPL*, and *PER* were computed across all phones and we can observe the improvement of global accuracy of automatic phonetic segmentation (for all analyzed criteria) when reduced pronunciation is used, i.e. both mean values and standard deviations decreased significantly. Significant improvement of segmentation accuracy about 3.2 ms in average (across all criteria *SPB*, *SPE*, *CLP*) was observed when reduced pronunciation was used instead of canonical one, slightly smaller improvement about 1.9 ms when the lexicon contained more pronunciations variants. Secondly, the results for phone categories are in Fig. 3, here we can see in more details achieved segmentation accuracy for phones of similar character [15]. Especially, the mean values and standard deviations for VOW, FRI, VOWNH, FRIAFF were significantly improved for criteria *SPB*, *SPE*.

Although presented results were obtained by experiments performed on small evaluation subset of manually segmented utterances, the results already proved the contribution of the information about pronunciation reduction. As currently further manual segmentation of the evaluation test utterances is under the processing we assume that these results will be precised during the presentation at the workshop.

	<i>canonical</i>	<i>reduced</i>	<i>variants</i>
<i>SPB</i> [ms]	-5.24 ± 47.59	-1.05 ± 18.96	-3.39 ± 39.53
<i>SPE</i> [ms]	-4.16 ± 41.39	-1.64 ± 22.49	-1.42 ± 38.53
<i>CPL</i> [ms]	-3.17 ± 20.75	-0.07 ± 21.65	-2.07 ± 19.12
<i>PER</i> [%]	21.04	1.18	14.94

Table 2. The overall results of phonetic segmentation for all phones**Fig. 3.** The results of phonetic segmentation for particular phone groups (red - canonical pronunciation, blue - reduced, magenta - more pronunciation variants)

7 Conclusions

This study is the first step in further detailed research on pronunciation reduction in NCCCz which is also supposed to be used for better modelling of spontaneous speech for recognition purposes. The realized pilot analysis of HMM-based phonetic segmentation accuracy with respect to the usage of canonical or reduced pronunciations is the main contribution of this paper. The experiments done with the speech from Nijmegen Corpus of Casual Czech (NCCCz) with very strong spontaneous nature demonstrated the significant impact of pronunciation reduction on the proper acoustic modelling of spontaneous speech (applied currently on phonetic segmentation). Further contribution of this paper is in the basic information about NCCCz corpus and basic its lexicon containing typical casual words. The created tool LexFix also represents the important contribution of this work because it supports generally lexicon editing with possible checks of a word context in the corpus same as listening selected. Finally, the created

evaluation database with utterances from NCCCz completed by manual labels at phonetic level is the last important contribution. It was used in the experimental part described in this paper but it is suitable for evaluation purposes in general and it is supposed to be used within further experiments with spontaneous or casual speech.

Acknowledgments. Research described in this paper was supported by the internal CTU grant SGS14/191/OHK3/3T/13. We would also like to thank Zdenek Patc and Helena Pollakova for their work done in manual phonetic segmentation. The work on the creation of NCCCz and its lexicon was funded by the European Young Investigator Award given to the fourth author.

References

1. Kohler K. J.: Segmental reduction in connected speech in German: Phonological facts and phonetic explanations. In W. J. Hardcastle, A. Marchal (Eds.), *Speech Production and Speech Modelling*. Kluwer Academic Publishers, pp. 69-92, 1990.
2. Ernestus M.: Voice assimilation and segment reduction in Dutch: A corpus-based study of the phonology-phonetics interface. Utrecht: LOT, 2000.
3. Johnson K.: Massive reduction in conversational American English. In K. Yoneyama, K. Maekawa (Eds.), *In proc. of the 10th International Symposium on Spontaneous Speech: Data and Analysis*. Tokyo, Japan, 2004.
4. Hinton G. and et al.: Deep neural networks for acoustic modeling in speech recognition, *Signal Processing Magazine*, 2012.
5. Vanek J., Psutka J.: Gender-dependent acoustic models fusion developed for automatic subtitling of parliament meetings broadcasted by the Czech TV. In *proc. of Text, Speech and Dialog*, Brno, Czech Republic, 2010.
6. Ernestus M., Kockova-Amortova L. and Pollak P.: *The Nijmegen Corpus of Casual Czech*. LREC 2014, 26-31 May, Reykjavik, Iceland, 2014.
7. Pollak P. and Cernocky J., *Czech SPEECON adult database*. Technical report, Nov. <http://www.speechdat.org/speecon>, 2003.
8. Pollak P., Cernocky J. and et al.: *Speechdat(E) - Eastern European telephone speech databases*, In *proc of XLDB*. Athens, Greece, 2000.
9. Siemund R., Hge H., Kunzmann S. and Marasek K.: *SPEECON - Speech data for consumer devices*. In *Proc. of the LREC 2000*, Athens, Greece, 2000.
10. Hanzl V., Pollak P.: *Tool for Czech Pronunciation Generation Combining Fixed Rules with Pronunciation Lexicon and Lexicon Management Tool*. In *proc. of LREC 2002*, Las Palmas de Gran Canaria, Spain, 2002.
11. Cmejla R., Ruz J., Bergl P. and Vokral J.: Bayesian changepoint detection for the automatic assessment of fluency and articulatory disorders. In *Speech Communication*, 55(1), 178-189, 2013.
12. Mizera P., Pollak P.: Accuracy of HMM-based phonetic segmentation using mono-phone or triphone acoustic model. In *proc. of Applied Electronics*, Pilsen, CR, 2013.
13. Schwarz P.: *Phoneme recognition based on long temporal context*, PhD Thesis, Brno University of Technology, 2009.
14. Povey D., Ghoshal A. and et al.: *The Kaldi Speech Recognition Toolkit*, In *proc. of ASRU*, Hawaii, USA, 2011.
15. Pollak P., Volin J., and Skarnitzl R.: *Phone Segmentation Tool with Integrated Pronunciation Lexicon and Czech Phonetically Labelled Reference Database*. In *proc of LREC*. Marrakech (Morocco), 2008.