Czech Spontaneous Speech Collection and Annotation: The Database of Technical Lectures

Josef Rajnoha and Petr Pollák

Dept. of Circuit Theory, Czech Technical University, Prague
{rajnoj1,pollak}@fel.cvut.cz

Abstract. Applying speech recognition into real working systems, spontaneous speech recognition has increasing importance. For the development purposes of such applications, the need of spontaneous speech database is evident both for general design or training and testing of such systems. This paper describes the collection of Czech spontaneous data recorded within technical lectures. It is supposed to be used as a material for the analysis of particular phenomena which appear within spontaneous speech but also as an extension material for training of spontaneous speech recognizers. Mainly the presence of spontaneous speech phenomena such as higher rate of non-speech events, changes in pronunciation, or sentence irregularities, should be the most important contribution of the collected database for the training purposes in comparison to the usage of available read speech databases only. Speech signals are captured in two different channels with slightly different quality and about 14 hours of speech from 15 different speakers are currently collected and annotated. The first analyses of spontaneous speech related effects in the collected data have been performed and the comparison with read speech databases is presented.

1 Introduction

Current Automatic Speech Recognition (ASR) systems are used much more frequently in the communication between a human and a machine. We can meet e.g. voice controlled device operation, dictation machines, or any general recognition of spoken speech for purposes of transcription of records, on-line TV subtitles, etc. The speech at the input of such systems becomes more and more natural and the ASR must deal with the effects of spontaneous talk. Consequently, it makes the recognition much more difficult [1].

The first issue which is usually met appears during the training of speech recognizers. On the acoustic level, read speech databases are usually used for the training of acoustic HMM models. As these databases are collected mainly with this special effort, they often contain phonetically balanced material. But the level of speech disfluencies and other non-speech events typical for spontaneous utterances is rather small within these databases. Moreover, additional non-verbal information comes with spontaneous speaking style, such as changes in intonation or gestures. The need of the presence of spontaneous utterances

A. Esposito and R. Vích (Eds.): Cross-Modal Analysis, LNAI 5641, pp. 377–385, 2009.

in the training databases is evident and the collection of spontaneous speech corpora is necessary for this purpose [2, 3]. It was also proved by our first experiments with small vocabulary digit recognizer that modelling of non-verbal events improved recognition accuracy [4]. To generalize this experiment for large vocabulary spontaneous speech recognition in the future, it is another purpose of this spontaneous speech collection.

As rather small amount of spontaneous speech data is currently available for Czech language from publicly available sources and as we have the opportunity of quite efficient collection of spontaneous speech data with similar topics and spontaneous speaking style, we have started collecting of Czech spontaneous speech database of technical lectures which is described in this paper. Strong effort is paid to the annotation of the collected data, especially from the point of view of precise description of non-speech events and other disturbing issues typical for spontaneous speech appearing in collected utterances.

2 DSP Lecture Collection

Our current collection consists of the recordings captured within Digital Signal Processing (DSP) lectures at our department, containing periodic doctoral reports and selected lectures of DSP-specialized courses. Each session concerns about 20–30 minutes of speech on signal processing theme, which is usually prepared in advance. The speech is then more fluent and better pronounced while the speaking style is still very spontaneous. The database involves single speaker utterances rarely disturbed by the question from the audience or by an expected answer to posed question.

Similar topics of the speech are very important for further testing of spontaneous speech recognition systems. This collection could extend currently available training databases which are important by involving phonetically balanced read utterances on the other hand. Currently, we have the first set of recordings which consists of 3 lectures and 32 reports. It gives about 14 hours of spontaneous speech from 15 different male speakers.

3 Recording Facility

Commercial wireless recording system was chosen to capture speech (see Figure 1). It gives the speaker a freedom of movement even in our case of two-channel recording. The system is designed to provide a high quality signal with respect to requested portability but also with the intention of possible connection to standard PC system. Remote control center provides necessary monitoring and easy adjusting the signal intensity in the case of possible saturation or low signal-level.

3.1 Hardware

Two-channel recording is performed by wireless transmission sets from Sennheiser. Each set consists of body pack transmitter, rack mountable receiver and omnidirectional lapel microphone (EW112 G2 set) or super-cardioid headset microphone

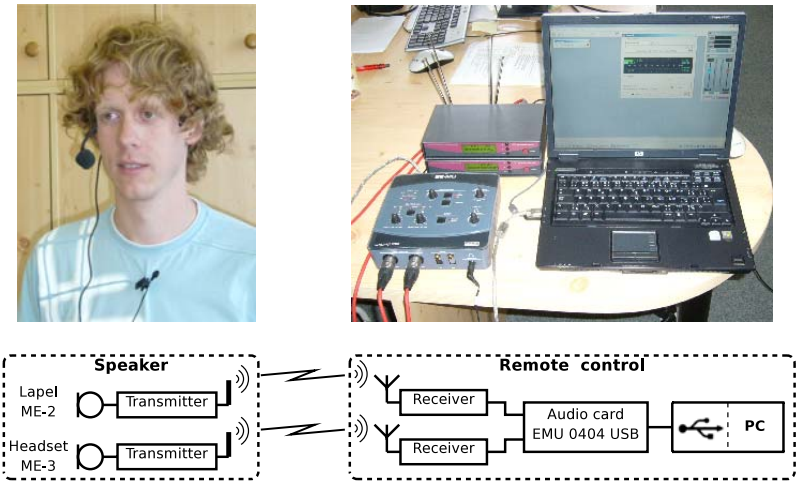

Fig. 1. Recording platform scheme with EW112 G2 and EW152 G2 wireless transmission system from Sennheiser

(EW152 G2). The lapel microphone is usually attached to the speaker's shirt about 10 cm from mouth. It captures higher level of background noise due to its omnidirectional receiving characteristics. The distance between the microphone and the speaker's mouth is also variable as the speaker turns his head to the table or presentation slides and to the audience. It causes audible fluctuations of signal intensity in the final record. The headset microphone is intended to capture high-quality speech signal with maximum movement and close proximity signal reproduction.

Received signal is digitized in dual-input USB sound card E-MU 0404 which enables direct sound-level manipulation for each channel and low-latency headphone monitoring. USB port provides possible connection to any standard PC for data storage and signal intensity monitoring without using additional tools.

We store the signal with maximum quality. The recording set enables to capture two-channel signal with 48 kHz sampling rate and with 16 bits precision per sample. This sampling frequency is suitable for integer-ratio resampling to the most frequently used 16 kHz sampling rate. The setting of sound-level was optimized within the first records to have balanced signal intensity with low appearance of saturation for loud events and audible signal for low speech level.

3.2 Software

The recording is performed using freely available software *WaveLab Lite* included in sound card package. It enables full dual channel recording support, i.e. data input monitoring and mastering (e.g. adjusting important points in the wave during recording) with minimum other superfluous functionalities.

On the other hand, no support is available from the point of view of speech database organization. Each further information, including signal segmentation and transcription is done later independently of the recording.

4 Signal Segmentation and Annotation

The presentations are recorded always at once and these several minutes long waveforms need to be segmented to shorter parts and annotated. The first recordings have been segmented and annotated fully manually but we suppose to use some automated or semi-automated procedure for the next collection for saving of hard manual work.

Freely distributed speech segmentation, labeling, and transcription software *Transcriber* (distributed by LDC [5]) is used for signal segmentation and manual dataset annotation. It gives full support for the first steps in speech corpus creation. The required steps of signal processing and annotation are described in the following sections.

4.1 Segmentation

Each utterance is divided into sentences, which are the most suitable form for the next processing and for training of ASR system. As the transcription is not known, the start- and end-points of the sentences are found manually.

As shown in [1] the start-point and end-point of the spontaneous sentences can be very different from fluent read speech. False starts, repairs, and repetitions can appear and the sentence can be also very long. Moreover, the speech can be silent at the end of the sentence. Important parts of speech can be "hidden" in the environmental noise, but it is important to keep this information in the correct segment. Regarding the mentioned effects, the signal is segmented with special effort to correct placement of the sentence boundaries. Short part of the signal with non-speech information is kept at the beginning of each sentence if it is possible. It could be useful in further noise compensating algorithms.

Longer segments without speech activity between two separate blocks of speech are cut and marked as pause segments. The original long recordings are kept in the database for the purposes of further research.

4.2 Orthographic Transcription

The speech content is transcribed in the form of orthographic annotation. As in other database projects ([3], [6]), standardized rules are used for the annotation. The transcription procedure is divided into several steps.

Only speech content is transcribed in this first annotation step and other effects are omitted. Punctuation is not noted as it is not important for training a phoneme model-based speech recogniser. A lower-case form is used for all characters.

The speech is rewritten in the form it is exactly spoken, even in the case of colloquial language or mathematical expressions, which are typically present in technical speech. Special transcription is used for spelling, mispronunciations and foreign words (see Table 1).

Small changes in pronunciation are not transcribed as they are supposed to be covered by the variability of the HMM modelling of elementary phonetic elements.

The content of collected utterances is not known in advance and as it can be quite unusual many times, some words may be often difficult to be recognized.

Table 1. Typical effects in spontaneous speech and their annotation

Event	Transcription
spelled sounds	**'$' prefix and correct pronunciation variant for given sound**
mispronunciations, small mistakes	**'*' prefix character**
strong mispronunciation	**'**' mark**
foreign words	**'∼' prefix character**

The annotation is therefore checked by another annotator more than once to guarantee the correct transcription as much as possible.

4.3 Phonetic Transcription

Phonetic transcription is not part of the annotation. It can be automatically generated by tool *transc* [7] in the next annotation phase. This tool generates the phonetic transcription from orthographic transcription on the basis of grapheme-to-phoneme conversion rules.

Exceptions in pronunciation are covered by special dictionary or by the annotation of special pronunciation as noted within orthographic transcription. More specific pronunciation irregularities can be also marked in the form (*word/ pronunciation*) which defines correct pronunciation variant for the given word.

4.4 Non-speech Event Annotation

Spontaneous speech differs strongly from read speech mainly by the fact that speakers often need to think about succeeding words. It causes much more silent pauses, lengthenings, or filled pauses in such speech. These effects are marked together with other environmental events in the next annotation step. As the receiving characteristics of particular microphones are different, also the transcription of non-speech events can differ slightly for particular channels.

Non-speech events are divided into several classes according to Table 2.

Table 2. Description of annotated non-speech events

Mark	Description
Speaker-generated events	
[*mlask*]	**lip smack**
[*dech*]	**breath**
[*fil*]	**filled pause**
[*smich*]	**laugh**
[*ehm*]	**throat clear**
[*kasel*]	**cough**

Mark	Description
Other speaker distortions	
[*cockt*]	**cocktail-party effect**
[*other*]	**other speaker**
Background noise	
[*sta*]	**stationary noise**
[*int*]	**non-stationary interruption**

- *Speaker-generated non-speech events* – As they are produced by the speaker the same way as speech, speaker-generated non-speech events always occur between words. They can be therefore annotated as another word, using a key word in square brackets (e.g. '*word1* [*fil*] *word2*'). Used speaker-generated events which appear typically in spontaneous speech are listed in Table 2.
- *Background noise events* – Even the speech is recorded in quiet environment, it can still contain disturbing noise which must be annotated. But environmental distortion can overlap particular words so special rules are used for better description of noise disturbance within the speech. If the noise appears only in the pause between words, it is marked similarly to speaker-generated events. When the following speech is affected, starting and ending mark is used, e.g. "*word1* [*int*−] *word2 word3* [−*int*] *word4*". This convention corresponds to rules used in transcription tool *Transcriber*.
- When "[*sta*]" mark is used without beginning and ending mark, it should be placed at the beginning of the utterance transcription and it means the presence of stationary noise in whole signal.
- *Other speaker in background* – As the speech is recorded within lectures, the audience present in the room can influence the speech and distortion from other speaker can be present. We resolve two different situations (see Table 2), either the distortion appears within speech pause ([*other*]) or more speakers are talking simultaneously ([*cockt*]).

5 Dataset Analysis

The final structure of the database was defined within the processing of the first data. It involves long recordings, cut segments with particular sentences, orthographic transcription of speech content with possible irregular pronunciation and non-speech event description. This section provides general comparison of collected spontaneous database with other available read speech databases. Even the amount of currently available spontaneous data is rather small, it describes main attributes of the collection and overall character of collected speech.

Currently transcribed part of the database contains 7830 different words from total amount of 63000 words. The spontaneous character of the speech is evident from the occurrence of colloquial and slang words in comparison to standard written corpora (4.6 % of colloquial words against e.g. 0.08 % of these words in Czech LC-Star2 [8]). The final set of words contains about 21.2 % of topic-related words.

5.1 The Speech Intelligibility

Table 3 presents the comparison of the amount of correct, mispronounced and unintelligible words in different speech corpora (the percentage is related to the amount of all words in the database). 'SPEECON' and 'CZKCC' are read speech databases, 'Lectures' is our spontaneous speech collection.

The occurrence rate of words with small mispronunciation in the spontaneous database is comparable to large read speech collections. But the rate of mispronounced words is higher. It is caused mainly by the effect of repetitions and

Table 3. Word distribution in particular databases

database	words	mispronunciations		unintelligible/incomplete words	
SPEECON	561 716	1157	(0.21 %)	1768	(0.31 %)
CZKCC	1 067 412	1689	(0.16 %)	1902	(0.18 %)
Lectures	63 000	85	(0.14 %)	445	(0.71 %)

repairs which interrupt speech in the middle of a word. On the other hand, despite the spontaneous character of the utterances in our new database, the rate of mispronunciations is still rather small and collected speech seems to be good material for training purposes.

5.2 Occurrence of Non-speech Events

The presence of non-speech events in training databases is important for robust ASR systems. Cleared read speech databases were compared in terms of non-speech event occurrence. Due to significant differences in speech content, the part of SPEECON database which contains spontaneous utterances was analysed separately from the read speech subset for this purpose.

Tables 4 and 5 show the amount of non-speech events marked by the human annotator (percentage is again related to the amount of words in the fragment of particular database). We use more precise description of non-speech events in our currently created database, but other databases use simpler categorization. Simplified two classes of speaker non-speech events (filled pause, other event) were therefore analyzed.

Table 4. Occurrence of filled pauses in inspected databases

database	words	filled pauses	
SPEECON read	146537	344	(0.23 %)
SPEECON spont.	34954	1512	(4.33 %)
CZKCC	244044	153	(0.06 %)
Lectures	54314	1449	(2.67 %)

It can be seen in Table 4 that spontaneous collections contain significantly higher rate of filled pauses than read utterances as it is typical for spontaneous speech [9]. On the other hand, spontaneous speech is more fluent and without longer pauses present during recordings of separated read utterances. They are frequently followed by lip smack and audible breath. The occurrence of other events is therefore lower for spontaneous speech (see Table 5).

The influence of recording conditions, mainly chosen microphones and their position, or further background environment also yield to different rates of non-speech events in compared databases. Moreover, this difference can be affected slightly by the annotation inconsistency [10]. Finally, it causes significant difference also between both read speech databases.

Table 5. Occurrence of other speaker-generated events in inspected databases before and after forced-alignment

database	words	other non-speech events			
		annotated		aligned	
SPEECON	181517	33125	(18.25 %)	29934	(16.49 %)
CZKCC	244044	15728	(6.44 %)	9941	(4.07 %)
Lectures	54314	203	(0.37 %)	134	(0.25 %)

Due to the fact mentioned above, it is reasonable to mark only significant non-speech events for modelling purposes. Forced-alignment commonly with reached acoustic score analysis was used to reduce the occurrence of inexpressive non-speech events. Table 5 shows the number of retained events for all 3 databases. Such corrected data are then supposed to represent better material for further training.

6 Conclusions

The paper presented the collection of Czech spontaneous database. The most important contributions are summarized in following points.

- Recording scenarios and recording platform for creation of Czech spontaneous database of lectures were defined and the first sessions were collected. Currently, the collection contains about 14 hours of spontaneous speech and the recording continues. The final structure of the database involves whole long recordings, segmented signal cut to particular sentences, commonly with orthographic transcription with precise annotation of non-speech events.
- Annotation conventions for orthographic transcription of spontaneous speech were designed and the first data were annotated. Extended set of non-speech events was defined to describe speaker-generated and environmental non-speech events more precisely.
- According to our assumption, the first analyses showed higher rate of slang and colloquial words in comparison to standard written corpora. We have observed approximately 21.2 % of topic-related and 4.3 % of colloquial words in the presented spontaneous collection. Also the rate of interrupted or unintelligible words is slightly higher in comparison to standard read speech collections. Nevertheless, the speech fluency is still high and the data are suitable for further usage.
- The re-alignment procedure decreased the amount of inexpressive non-speech events in the collected data. Consequently, more precise modelling of non-speech events is supposed to be achieved when these data are used for training purposes.
- Our preliminary experiments on filled-pause recognition showed significant contribution of proposed spontaneous database in non-speech event recognition task. Using our database for training, the insertion error rate decreased

by approx. 80% against the case with read speech training data. Having spontaneous speech data from presented database, the full application of non-speech event modelling into spontaneous speech recognizer can be now the next step of our activities.

Acknowledgements

The research was supported by grants GAČR 102/08/0707 "Speech Recognition under Real-World Conditions", GAČR 102/08/H008 "Analysis and modelling biomedical and speech signals", and by research activity MSM 6840770014 "Perspective Informative and Communications Technicalities Research".

References

[1] Shriberg, E.: Spontaneous speech: How people really talk, and why engineers should care. In: Proc. Eurospeech 2005, Lisbon, Portugal, pp. 1781–1784 (2005)
[2] Trancoso, I., Nunes, R., Neves, L., Viana, C., Moniz, H., Caseiro, D., Mata, A.I.: Recognition of classroom lectures in european Portuguese. In: Proc. Interspeech 2006, Pittsburgh, USA (2006)
[3] Psutka, J., Radová, V., Müller, L., Matoušek, J., Ircing, P., Graff, D.: Large broadcast news and read speech corpora of spoken Czech. In: Proc. Eurpospeech 2001, Ålborg, Denmark, pp. 2067–2070 (2001)
[4] Rajnoha, J., Pollák, P.: Modelling of speaker non-speech events in robust speech recognition. In: Proceedings of the 16th Czech-German Workshop on Speech Processing, Academy of Sciences of the Czech Republic, Institute of Radioengineering and Electronics, Prague, pp. 149–155 (2006)
[5] Barras, C., Geoffrois, E., Wu, Z., Liberman, M.: Transcriber: A free tool for segmenting, labeling and transcribing speech. In: Proc. of the First international conference on language resources & evaluation (LREC), Granada, Spain, pp. 1373–1376 (1998)
[6] Pollák, P., Černocký, J.: Czech SPEECON adult database (November 2003), `http://www.speechdat.org/speecon`
[7] Pollák, P., Hanžl, V.: Tool for Czech pronunciation generation combining fixed rules with pronunciation lexicon and lexicon management tool. In: Proc. of LREC 2002, Third International Conference on Language Resources and Evaluation, Las Palmas, Spain (May 2002)
[8] LC-STAR II project site, `http://www.lc-star.org/`
[9] Gajić, B., Markhus, V., Pettersen, S.G., Johnsen, M.H.: Automatic recognition of spontaneously dictated medical records for Norwegian. In: COST 278 and ISCA Tutorial and Research Workshop - ROBUST 2004 (2004)
[10] Rajnoha, J.: Speaker non-speech event recognition with standard speech datasets. Acta Polytechnica 47(4-5), 107–111 (2008)