Multi-Channel Database of Spontaneous Czech with Synchronization of Channels Recorded by Independent Devices

Petr Pollák, Josef Rajnoha

Czech Technical University in Prague, Faculty of Electrical engineering
Technick 2, 166 27 Praha 6, Czech Republic
pollak@fel.cvut.cz, rajnojos@fel.cvut.cz

Abstract

This paper describes Czech spontaneous speech database of lectures collected at Czech Technical University in Prague, commonly with the procedure of its recording and annotation. In this article, special attention is paid to the description of time synchronizations of signals recorded by two independent devices. This synchronization is based on cross-correlation analysis with simple automated selection of suitable short signal subparts. The database contains 21.7 hours of speech material recorded in 4 channels with 3 principally different microphones. The annotation of the database is composed from basic time segmentation, orthographic transcription, pronunciation lexicon, session and speaker information, and the documentation. The collection and annotation of this database is complete and its availability via ELRA is currently under preparation.

1. Introduction

As current applications of Automatic Speech Recognition (ASR) work normally with recognition under natural speaking conditions, the need for spontaneous or generally informal speech recognition increases. Such speech has a principally different nature in comparison to read speech (Shriberg, 2005), it mainly contains a higher amount of speaker non-speech events such as hesitations, repetitions, broken sentences, colloquial words, etc. On the basis of this fact, the availability of less formal or spontaneous speech data started having great importance for the research in the field of speech recognition within recent years.

Speech databases which are used most frequently for training and evaluation of ASR usually contain read speech, e.g. (SpeechDat, 2010), however, there is already the big group of spontaneous speech databases; Switchboard, CALLHOME, or generally other conversational telephone or broadcast speech material. But these databases are available mainly for English or for other important world languages. The first rather spontaneous Czech speech database appeared in LDC catalogue in the second half of last year (Kolář et al., 2009) and (Kolář and Švec, 2009) and it contained broadcast conversation from a talk show at Czech Radio 1. Generally, the amount of publicly available Czech spontaneous speech data is rather small, so this was the basic motivation for the collection of the presented database which was started two years ago. During the collection of this database, Nijmegen Corpus of Casual Czech also started being created (Kočková-Amortová et al., 2010) but this database is not publicly available yet.

In this paper, we describe the collection and annotation of the newly created database of technical lectures in the Czech language, commonly with solutions of other related technical problems. It is very difficult to collect absolutely spontaneous speech, e.g. described in (Obuchi and Amano, 2007), so the possibility of recordings of regularly scheduled lectures was a good opportunity to have spontaneous speech data with rather easy and efficient collection. As speakers were supposed to think about the expression of the idea roughly prepared in advance, not about the idea itself, collected speech is rather fluent and better pronounced, but we can also suppose the appearance of phenomena typical for spontaneous speech in this data. Our collection was inspired also by similar collections done for other languages, e.g. for Portuguese described in (Trancoso et al., 2006) or for English within the MIT spoken lecture processing project (Glass et al., 2005).

The first basic description of this collection was presented in (Rajnoha and Pollák, 2009). In this paper, we will describe final content, slightly modified recording setup for new sessions, and annotation rules of recorded data of the complete database. As two additional channels were recorded by the independent device, we had to solve several technical problems, mainly time-synchronization of collected data whose description is supposed to be the second important contribution of this paper.

2. Database recording setup

The described corpus comprises recordings captured within Digital Signal Processing (DSP) lectures at the Czech Technical University in Prague, containing periodic doctoral reports in the field of speech and biological signal processing and selected lectures from the DSP course. The decision about the collection of this data was made on the basis of the above mentioned need of spontaneous data but also on the basis of very similar topics discussed at these lectures.

As it is usual for similar data collections, we decided to record the data from several channels of different quality at once within one session to increase efficiency of the collection. In the beginning, we started recording with two synchronous channels using one super-cardioid headset microphone EW152 G2 from Sennheiser which was supposed to collect a high quality signal and with one omni-directional lapel microphone Sennheiser EW112 G2 collecting a higher level of background noise. Both microphones were used with wireless transmission sets with mountable body pack transmitters, rack mountable receivers, and dual-input USB sound card E-MU 0404. It

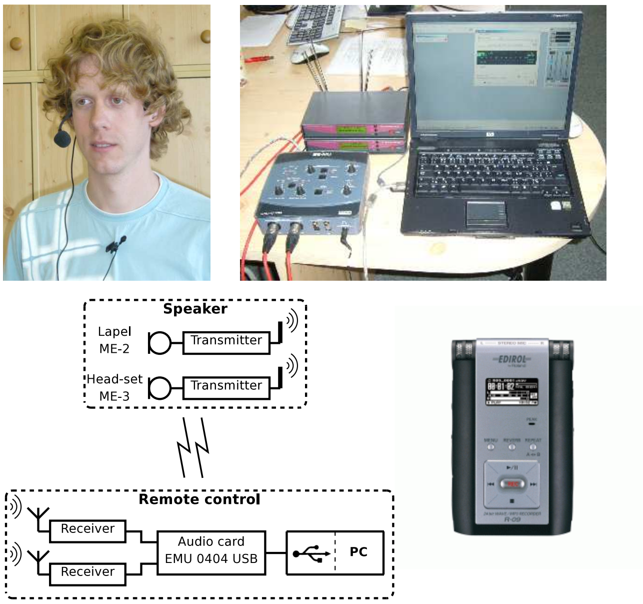

Figure 1: Recording platform: block scheme and illustrative photos of microphones EW112 G2 and EW152 G2, wireless transmission systems, and Edirol R09 recorder

enabled direct sound-level manipulation for each channel and low-latency headphone monitoring. The block scheme of recording platform commonly with illustrative photos is in the fig. 1.

The second half of the sessions also contains an independent stereo signal collected by Edirol R09 recorder, an affordable commercial device from Roland for the recording of audio data. This device allows direct digitalization of collected data at adjustable sampling frequency and its storage as MP3 audio or without any compression as standard 16-bit linear PCM, see illustrative photo on the fig. 1. Though it is possible to use external microphones with this recorder, we have used built-in stereo microphones in our recordings to simulate the collection with table-top microphone. The recorder was placed on the table approximately 1-2 m from the speaker. The distance was not fixed as speakers were often moving during the presentations.

finally, wide-band speech was recorded in each channel, i.e. sampling frequency of recorded signals was 48 kHz so integer-ratio down-sampling to 16 kHz, the most frequently used sampling frequency in speech technology systems, is now easily possible.

Recordings were performed in a standard lab room with an area of about 50 m^2. Concerning the background environment, the recording conditions were rather quiet. Sometimes, the windows in the room were opened, so a rather minor level of background street noise could be present. Mainly the head-set microphone was supposed to record the signal with very low level background noise, on the other hand, signals recorded by lapel microphone or Edirol R09 contain higher level of background noise.

The SNR was estimated for all signals in DB and its distributions for particular channels are shown in fig. 2. Concerning SNR estimation, the power of background noise was estimated as the average of 5% of the smallest short-time powers,the power of speech was then estimated from resting 95% of the short-time powers.

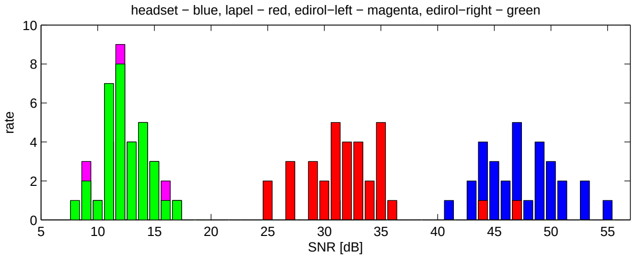

Figure 2: Histograms of SNRs in particular channels

3. Synchronization of independent multi-channel recordings

As it was mentioned above, the second half of the database contains signals recorded by two independent devices for particular channels. Consequently it means, that the synchronization of each channel is not guaranteed automatically. Generally two problems are met in such situation. Firstly, though the same sampling frequency is set, minor differences in its value can always be observed for any two independent devices. It can cause asynchronous end of multi-channel signal, especially when recorded sessions are rather long as it is in our case. Maximal measured difference in sampling rates in this collection was only 0.03%, typical value was less then 0.002%. As it was less then 1 Hz for 48 000 Hz sampling rate, it was an acceptable variation. So finally, only the time-synchronization of the independently recorded signals had to be solved in our case because when two independent devices are used they cannot be turned on exactly at the same time.

3.1. Synchronization procedure

In several recording setups we can find the synchronization based on artificial reference sound (beep), e.g. the synchronization of speech transmitted via GSM with in-car platform speech within the SpeechDat-Car project (van den Heuvel et al., 1999). As such a solution is not possible in the recording of fluent and spontaneous speech, we have used the standard technique based on cross-correlation analysis between non-synchronous channels. It is rather well known principle which was used also by other authors with more or less different setup, e.g. in (Brandstein and Silverman, 1997). As we are working with quite long signals, the parameters of the synchronization procedure must be set to avoid an enormous increase of computational costs. Finally, our simple and robust correlation-based synchronization procedure of independently recorded channels was based on the following particular steps.

Step 1: Manual rough pre-synchronization

Though the first step is manual, it represents quite standard procedure after any recording of any session. Within the feedback about quality of recorded speech the channels are roughly synchronized. The beginnings of signals for high quality channels were set definitely, for non-synchronous channels we always kept a small reserve before the utterance beginnings. Consequently, the channels were nearly synchronous, i.e. the delay between the signals should have not been grater than 1-2 s. Precise synchronization was then performed automatically in next steps.

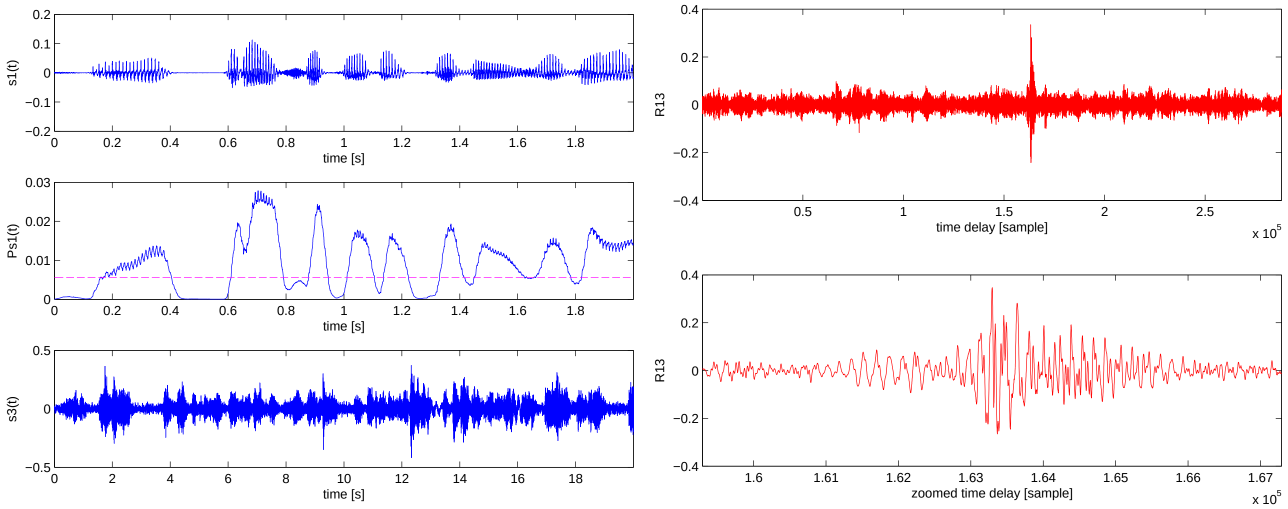

Figure 3: Illustration of synchronization procedures: signal frames, instantaneous power, cross-correlation

Step 2: Selection of short-time frame $s_1(t)$
The basic short-time speech frame $s_1(t)$ from the high quality channel (headset) were selected. The length of this frame should be rather high to contain a unique speech sequence, on the other hand, increasing length means higher computational costs. Finally, we used the frame of length 2 s.

Moreover, additional properties of the chosen frame were analyzed. To avoid the selection of noise (pause) frame, the minimal power threshold was set. The selected frame should have also contained a minimal amount of different rhythmic subparts. Boundaries of these subparts were obtained on the basis of $P_{s1}(t)$, i.e. instantaneous power of $s_1(t)$, with respect to the threshold set on 20% of achieved $P_{s1}(t)$ dynamics, see fig. 3.

Step 3: Selection of longer frame $s_3(t)$ or $s_4(t)$
This speech segment should be located approximately at the same part as $s_1(t)$ its duration should be longer. We worked with 20 s frame length for $s_3(t)$ or $s_4(t)$.

Step 4: Cross-correlation based delay detection
The estimation of delay between analyzed signals is then based on looking for the maximum of cross-correlation function between signals $s_1(t)$ and $s_3(t)$ computed by moving $s_1(t)$ over $s_3(t)$, i.e.

$$\tau_D = \underset{\tau}{\max \arg} \tag{1}$$

$$R_{13}(\tau) = \sum_{t=o}^{T} s_1(t) s_3(t+\tau) \tag{2}$$

where t represents discrete time, τ discrete time-delay, and T length of $s_1(t)$.

This computation was done in two steps. Firstly, the cross-correlation was computed roughly with the frame moving with higher step to save computational costs. Secondly, it was repeated with maximal precision within a small region of the found maximum in the first step, see fig. 3 for details.

Step 5: Repetition for other positions of $s_1(t)$
Above mentioned procedure was then repeated for five positions of $s_1(t)$ frame within the whole session. The analysis of differences between particular delays for all of these positions proved to us the information about sufficient accuracy of found delays and also about possible differences in sampling rates in particular devices.

3.2. Implementation and analysis of synchronization accuracy

The above described procedure worked well with sufficient accuracy. The correctness of the synchronization was also checked manually by listening tests. No audible echo (due to possible error of the synchronization) was observed. The time needed for this synchronization was rather small, although the implementation had been done in MATLAB which did not offer the fastest solution. Processing time needed for the synchronization of one session was very short, it was below 1 minute. In comparison to the duration of whole session it represented a typically real-time factor strongly below 0.1.

As two synchronous high quality signals and two synchronous signals from Edirol are available, the computation of delays can also be done on the basis of other signal couples. The choice of the signal from the lapel microphone is not suitable as the noise level in this channel is much higher and it means increasing estimation error. On the other hand, both channels from Edirol contain strongly correlated data, so the estimated delay can be checked by the computation from another couple of signals.

4. Description of Lecture Database

As mentioned above, collected data contains fluent utterances with topics from the field of digital signal processing. Particular features of collected data are summarized in the following paragraphs.

4.1. General description

General characteristics of the final database are in table 1. Each recorded session is kept in one long signal file with associated transcription. Further cutting on the bases of annotated boundaries is possible.

Signals and annotation files are organized in a simple structure as described in table 2, with separate directories for particular blocks of 10 sessions with additional directories containing the documentation and lexicon.

Corpus identification	**CSDSP10**
Total size	21.3 GB
Total length of speech	21.7 hours
Number of sessions	63
Average session duration	20 min
Number of speakers	22 21 males 1 female
Age of speakers	23-40
2 channel recordings 4 channel recordings (Edirol)	SES000 - SES029 SES030 - SES062

Table 1: General description of CSDSP10 database

Directories	Files
/	readme.txt
csdsp10/	
|- block00/ |- block01/ |- . . . |- block06/	*.wav *.trs *.snr
|- table/	lexicon.utf
|- doc/	csdsp10.pdf

Table 2: Directory structure of CSDSP10 database

4.2. SNR of particular channels

Average values of estimated SNR in particular channels are in table 3. We can see the really high quality headset channel in comparison to rather noisy Edirol (top-table) channels.

Channel ID	**SNR [dB]**
0 (headset)	46.45
1 (lapel)	32.18
2 (edirol-left)	12.61
3 (edirol-right)	12.36

Table 3: Average SNRs in particular channels

4.3. Database annotation

The annotation of collected speech signals has been done by Transcriber software (Barras et al., 1998) and (Boudahmane et al., 2010) and it contains orthographic transcription with marks for several non-speech events and segmentation into semantically consistent subparts.

4.3.1. Segmentation of long utterances

The typical length of a semantically consistent utterance subpart (segment) is approximately 5 s. The boundaries were placed carefully with the attention not to cut any word. When sufficient space between two words was not found, the length of a particular segment could be longer.

To save unnecessary manual work with settings of segment boundaries, rough segmentation of long recordings on the basis of VAD detection was done (Pollák and Rajnoha, 2009) and the found segment boundaries were later only corrected during annotation, commonly with the creation of orthographic transcription.

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

Table 4: Typical effects in spontaneous speech and their annotation

Mark	**Description**
	Speaker-generated events
$[mlask]$ $[dech]$ $[fil]$ $[smich]$ $[ehm]$ $[kasel]$	lip smack breath filled pause laugh throat clear cough
	Other speaker distortions
$[cockt]$ $[other]$	cocktail-party effect other speaker
	Background noise
$[sta]$ $[int]$	stationary noise non-stationary interruption

Table 5: Description of annotated non-speech events

4.3.2. Orthographic transcription

Utterances content has been annotated in the form of orthographic transcription. As in other database projects (SpeechDat, 2010), generally known rules have been adopted for the annotation. Lower-case form was used for all words, punctuation was not marked, speech was rewritten in the form as it was exactly spoken, including colloquial language or mathematical expressions which are often presented in recorded informal technical presentations.

Special conventions were used for the transcription of spelling, mispronunciations, or foreign words, see Table 4.

A special problem which had to be solved was related to the annotation of common words which appeared in this less formal speech. In accordance with the mentioned related collection of Nijmegen Corpus of Casual Czech (Kočková-Amortová et al., 2010) we tried to preserve regular forms of common words and possible variants of these words were solved at the level of pronunciation. On the other hand, extremely strange or strongly irregular pronunciations were transcribed using a special event mark.

Also the annotation of non-speech events was realized. Standard speaker non-speech events such as filled pause, breath, lip smack, cough, or laugh were annotated similarly as environment based non-speech events. All non-speech events were divided into classes according to the table 5.

The transcription is currently done only on the basis of the first channel, i.e. high quality headset microphone. Orthographic transcription is supposed to be the same for all recorded channels. Possible transcriptions of other channels can differ in non-speech event marks due to different levels of background environment picked-up by different microphones, but this transcription is not available yet.

The format of transcription files is hypertext XML supported by Transcriber software (Barras et al., 1998). All files containing Czech characters use UTF-8 encoding.

4.3.3. Quality checks

To achieve maximal accuracy of transcriptions, two level checks were realized. In the first step, a generalized spell-check procedure was applied. It was based on the combination of ispell tool usage while also looking for unknown words in already existing speech database transcriptions or other available lexica. In the second step, potentially strange transcriptions with new word forms were checked manually by another annotator, including the listening tests if it was necessary.

4.3.4. Additional annotation

The pronunciation of particular words was not transcribed at the level of each utterance. Pronunciation lexicon was created and it is now available as a standard supplement of orthographic transcription. Some words in this lexicon have multiple pronunciations to cover special cases of pronunciation variability. It appears mainly in the cases of words with foreign origin or other neologisms where the pronunciation is not standardized yet.

As the last part of database annotation, the information about quality of recorded signals (SNRs) as well as speaker code is stored for each session.

5. Conclusions

In this paper we have presented a newly created Czech language spontaneous speech database, along with some experiences from the collection and further processing of collected data. The most important contributions can be summarized as follows.

- New database of Czech spontaneous speech was created. It contains a large amount of data (more than 20 hours) collected by several channels with microphones of different quality. The database will be publicly available. The way of its distribution via ELRA is currently under negotiation.
- Collected data was precisely annotated on orthographic level with special focus to annotation of phenomena appearing in spontaneous or informal speech such as speaker and environmental non-speech events, annotation of common or colloquial words, etc.
- Long utterances were segmented into short semantically consistent segments and above mentioned transcription was done for these small segments.
- Automated synchronization of recordings from independent devices based on correlation analysis were proposed and used for signals in the presented database. Very good accuracy with minimal needs of manual interventions and with rather small computational costs was achieved for the solution of this task.

6. 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".

7. References

C. Barras, E. Geoffrois, Z. Wu, and M. Liberman. 1998. Transcriber: A free tool for segmenting, labeling and transcribing speech. In *Proc. of the First international conference on language resources & evaluation (LREC)*, pages 1373–1376, Granada, Spain.

K. Boudahmane, M. Manta, F. Antoine, S. Galliano, and C. Barras. 2010. Transcriber. A tool for segmenting, labeling and transcribing speech. online: http://trans.sourceforge.net.

M. S. Brandstein and H. F. Silverman. 1997. A robust method for speech signal time-delay estimation in reverberant rooms. In *Proc. ICASSP*, pages 375–378, Munich, Germany.

J. R. Glass, T. J. Hazen, D. Scott Cyphers, K. Schutte, and A. Park. 2005. The MIT spoken lecture processing project. In *Proc. of HLT/EMNLP on Interactive Demonstrations*, pages 28 – 29, Vancouver, British Columbia, Canada.

J. Kolář and J. Švec. 2009. Czech broadcast conversation MDE transcripts. Linguistic Data Consortium, Philadelphia.

J. Kolář, J. Švec, and J. Psutka. 2009. Czech broadcast conversation speech. Linguistic Data Consortium, Philadelphia.

L. Kočková-Amortová, P. Pollák, J. Rajnoha, and M. Ernestus. 2010. The Nijmegen corpus of casual Czech. To be submitted to Language Resources and Evaluation, Springer Netherlands.

Y. Obuchi and A. Amano. 2007. Always listening to you: Creating exhaustive audio database in home environments. In *Proc. of InterSpeech 2007*, pages 566–569, Antwerp, Belgium.

P. Pollák and J. Rajnoha. 2009. Long recording segmentation based on simple power voice activity detection with adaptive threshold and post-processing. In *Proc. of SPECOM 2009*, pages 55–60, St. Petersburg, Russia.

J. Rajnoha and P. Pollák. 2009. Czech spontaneous speech collection and annotation: The database of technical lectures. *Lecture Notes in Artificial Intelligence*, 5641(2009931057):377–385.

E. Shriberg. 2005. Spontaneous speech: How people really talk, and why engineers should care. In *Proc. Eurospeech 2005*, pages 1781–1784, Lisbon, Portugal.

SpeechDat. 2010. Web-page: Projects of SpeechDat family. on-line: http://www.speechdat.org/.

I. Trancoso, R. Nunes, L. Neves, C. Viana, H. Moniz, D. Caseiro, and A. I. Mata. 2006. Recognition of classroom lectures in European Portuguese. In *Proc. Interspeech 2006*, Pittsburgh, USA.

H. van den Heuvel, A. Bonafonte, J. Boudy, S. Dufour, P. Lockwood, A. Moreno, and G. Richard. 1999. SpeechDat-Car: Towards a collection of speech databases for automotive environments. In *Proc. of Nokia-COST249 Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, pages 135–13, Tampere, Finland.