Recording and Annotation of the Czech Speech Corpus

Vlasta Radová and Josef Psutka

University of West Bohemia, Department of Cybernetics,
Univerzitní 22, 306 14 Plzeň, Czech Republic
radova@kky.zcu.cz, psutka@kky.zcu.cz

Abstract. The paper reassumes our papers presented at the previous TSD workshops ([2], [3]) and concerns the Czech speech corpus which is being developed at the Department of Cybernetics, University of West Bohemia in Pilsen. It describes procedures of the corpus recording and annotation.

1 Introduction

The Czech speech corpus described in this paper is being developed at the Department of Cybernetics at the University of West Bohemia in Pilsen since 1998. The purpose of the corpus is to provide enough speech data for training of Czech continuous speech recognition systems. Our goal is to collect the corpus containing speech of at least 100 speakers, each of them is asked to read a set of 150 sentences. The texts to be read are selected from 3 Czech newspapers and have to satisfy several requirements that were specified in [3].

For the recording of the corpus a special program was developed that allows to record each utterance by two different microphones simultaneously. The function as well as the interface of the recording program is described in Section 2.

Before the utterances can be used for training of a speech recognition system they have to be annotated. The annotation is a process during that the speech data are divided into segments and various non-speech events (like lip smack, throat clear etc.) and various kinds of noise (e.g. chair squeak, door slam) are indicated in the utterances. In our corpus, the segments are defined as such parts of the utterance that contain either a whole sentence or some of the specified non-speech events. Since the non-speech events can be either isolated (it means no speech co-occurs with such an event) or they can co-occur with the speech, several types of descriptors were used to mark the place where the events occurred in the speech. The process of the corpus annotation is described in Section 3.

2 Recording of the Corpus

The corpus is recorded in an office room where no person but the speaker is present during the recording. However there is some noise from neighbouring offices in the recording room.

In order to record the corpus we developed a special recording program *Vlny* that allows to record the corpus sentence by sentence. Each sentence is recorded by two microphones simultaneously. One of the microphones is a close-talking microphone yielding utterances of a very high quality. The other one is a desk microphone that records utterances with a common office noise. Such an arrangement enables to yield two identical utterances that differs only in the amount of noise that they contain.

At the beginning of the recording session each speaker obtains some instructions. The instructions provide a general introduction to the project and a description of the task. The speakers are asked to read each sentence over before they utter and record it. They are also instructed to speak naturally and clearly in their usual accent at normal loudness. Each speaker is given a short demonstration how to operate the recording program.

The actual recording process consists of several steps, each co-ordinated by clicking on screen buttons. A typical window that appears after a sentence is recorded is given in Figure 1. At the top of the window there is the sentence that has just been read. Below it there are two windows each of them contains a figure of the signal from a microphone. The left window is for the close-talking microphone, the right window is for the desk microphone. The speaker has the possibility to check that the recording from any microphone is satisfactory by clicking on the corresponding *Play* button. Below the signal window there is also a scale window for each microphone. It shows the level of the signal at the input to the computer. The optimal level is approximately 50% and it should be adjusted by a supervisor during the demonstration phase, using a pre-amplifier for each microphone separately. During the actual recording it is not allowed to change the settings of the pre-amplifiers any more. However when the level of the signal in any microphone decreases below 10% or increases above 90% the speaker is asked by the program to read the sentence once more in a more or less

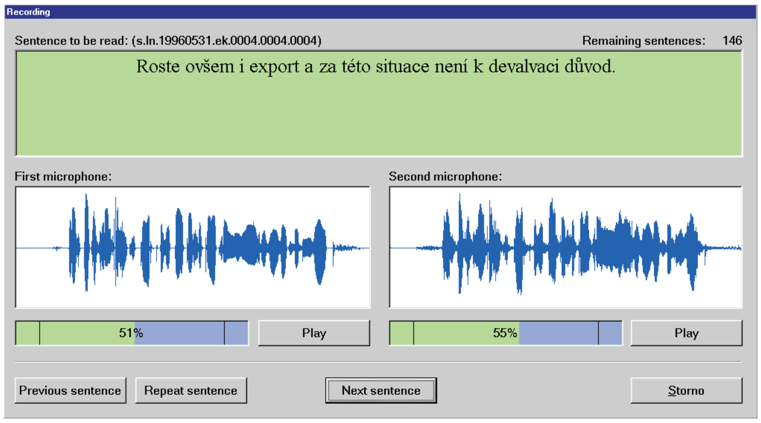

Fig. 1. A typical window of the recording program *Vlny*.

loud voice respectively. Similarly, when either the silence between clicking on the *Start* button and the beginning of the utterance, or the silence between the end of an utterance and clicking on the *Stop* button is shorter than 0.5 s the speaker is asked to repeat the sentence once more. Besides this, the speaker has a possibility to repeat the sentence on the base of his/her own decision by clicking on the *Repeat sentence* button. The *Previous sentence* button gives to the speakers a chance to go back to the previous sentence, the *Next sentence* button moves the speaker to the next sentence. The actual recording of a sentence is controlled by the *Start* and *Stop* buttons that appear after the *Next sentence* button is pressed.

3 Annotation of the Corpus

During the annotation process each utterance is transcribed in the form how it was really pronounced. It means with mispronunciation, unintelligible pronunciation, various non-speech events and various kinds of noise, if they occurred during the utterance. The rules that we use for the annotation of the corpus are as follows:

- Non-speech events and noises are indicated by a descriptor enclosed in square brackets. The descriptors contain only capitalized alphabetic characters and underscores and are drawn from the list in Table 1.

AH	LOUD_BREATH
COUGH	PAPER_RUSTLE
DOOR_SLAM	SIGH
GRUNT	TONGUE_CLICK
LIP_SMACK	UNINTELLIGIBLE
MM	MOUSE_CLICK
PHONE_RING	MIKE_OVERLOAD
THROAT_CLEAR	REMOTE_ENGINE
UH	NOISE
UM	KNOCK_ON_MIC
CHAIR_SQUEAK	MUSIC
CROSS_TALK	BACKGROUND_MUSIC
ER	SIGNAL_MISSING
LAUGHTER	SILENCE

Table 1. List of the descriptors of non-speech events and kinds of noise.

- The descriptor is placed at the point at which the non-speech event occurred. E.g.: *Sdělil to ministr školství [THROAT_CLEAR] Jan Sokol.*
- If a non-speech event overlaps a spoken lexical item the descriptor is placed close to the item that co-occurred with and the character "<" or ">" is

appended to the descriptor depending on whether the description is placed right or left of the co-occurring lexical item.
E.g.: *Akcie Komerční banky [CHAIR_SQUEAK>] poměrně zřetelně oslabily.*
Or: *Akcie Komerční banky poměrně [<CHAIR_SQUEAK] zřetelně oslabily.*
Both alternatives are equivalent and mean that a chair squeaked during the pronouncing of the word *poměrně*.

- If a non-speech event overlaps with more than one lexical item the character "/" is appended to the descriptor and the descriptor is then used like brackets bounding the lexical items.
E.g.: *Nabídka [NOISE/] udržela ceny ustálené. [/NOISE]*
- If the waveform is truncated (e.g. due to a recording error by the system) the symbol "~" is used to mark the incompletely spoken sentence:
E.g.: *Společnost bude z rozhodnutí vlády nejprve ~*
- Mispronounced but intelligible words are bounded with the mark "*". For example if the prompt was *Předsedové stran se domluvili* and the speaker has read *Předsedové stran se domlouvili* the utterance is transcribed as _Předsedové stran se *domlouvili*._
- Unintelligible words are replaced by the non-speech event [UNINTELLIGIBLE].

For annotation of all utterances we use *Transcriber*, a very useful tool for segmenting, labelling and transcribing speech [1]. It was developed in France and is freely available at `http://www.etca.fr/CTA/gip/Projets/Transcriber/`. A typical window of that tool is given in Figure 2. The upper part of the window is basically a text editor that allows to type the transcription of the utterances.

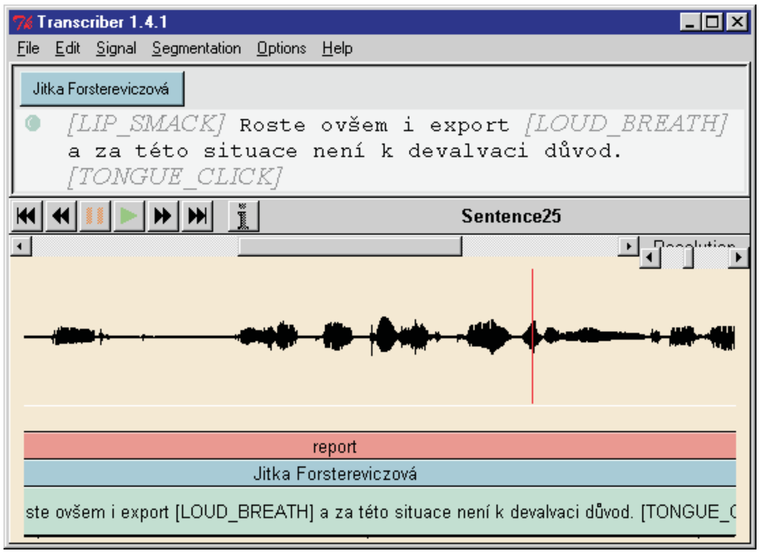

Fig. 2. A typical window of the *Transcriber*.

Besides this the speaker of the utterance can be specified here. At the bottom part of the window the speech signal is depicted.

4 Conclusion

The paper deals with the Czech speech corpus that is being developed at the Department of Cybernetics of the University of West Bohemia in Pilsen. It reassumes our papers presented at previous TSD workshops, where some questions about the corpus design were discussed [2] and algorithms for sentences selection [3] were described. The presented paper concerns with two next problems – utterances recording and corpus annotation.

5 Acknowledgements

The work was supported by the Ministry of Education of the Czech Republic, project no. VS97159, and by the Grant Agency of the Czech Republic, project no. 102/96/K087.

References

1. Barras C. at all: Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech. In: First International Conference on Language Resources and Evaluation (LREC) (1998)
2. Radová, V.: Design of the Czech Speech Corpus for Speech Recognition Applications with a Large Vocabulary. In: Sojka, P., Matoušek, V., Pala, K., Kopeček, I. (eds.): Text, Speech, Dialogue. Proc. of the First Workshop on Text, Speech, Dialogue. Brno, Czech Republic (1998) 299–304
3. Radová, V., Vopálka, P.: Methods of Sentences Selection for Read-Speech Corpus Design. In: Matoušek, V. at all (eds.): Text, Speech and Dialogue. Proc. of the Second Workshop on Text, Speech, Dialogue. Springer-Verlag, Berlin Heidelberg (1999) 165–170