CORPUS RECORDING AND CHECKING ON THE RECORDED DATA

Martin GRŮBER [1], **Milan LEGÁT** [2], **Daniel TIHELKA** [3]

Abstract: This paper describes a part of speech corpus recording procedure. It focuses on the supervision during the corpus recording and the checking on the recorded data. Several modifications of used software are presented.

Keywords: Speech synthesis, speech corpus, corpus recording.

1 INTRODUCTION

Concatenative speech synthesis is nowadays one of the most popular and most progressive method to produce synthetic speech. This method consists in building a database of speech units (so-called unit inventory) from natural speech and follow-up concatenation of these units into the correct form in order to produce the desired result.

For high-quality concatenative speech synthesis it is needed to be provided with a large speech corpus. Only in that case we have large database of natural speech and subsequently during the synthesis we have extensive line of realizations of different units. The corpus should be recorded by an experienced speaker under excellent conditions using top technical equipment. The speaker needs to be able to keep his speaking style constant during the whole period of recording.

Meeting all requirements for recording is very hard to please. This work deals with the last problem mentioned above, i.e. to lead the speaker to speak the same speaking style, the same speech power and to keep other possible conditions at the same level during all recording session. In addition we tried to reveal any abnormalities in both speech signal (e.g. any undesirable noise) and also glottal signal (e.g. battery discharge in laryngograph) which is very important for speech synthesis too.

2 SOFTWARE FOR RECORDING

For recording of our speech corpus, we used software which is specialized for this purpose. This software enables us to make and incorporate our own modules for recording, so we can guide the recording process.

2.1 Basic interface

First of all we had to create an interface that would be understandable for the speaker. He is the one, who communicates with the computer system during the recording. However, the speaker is not supposed to be familiar with computers, so the software has to be easy to control.

In fig. 1 you can see the same window as the speaker is presented with during the recording process. At the top of the window, there is text to be read, at the bottom there are control buttons. It is evident that it is easy to understand what to do and how to control the software.

[1]Ing. Martin Grůber, University of West Bohemia in Pilsen, Faculty of Applied Sciences, Department of Cybernetics, Univerzitní 22, 306 14 Pilsen, tel.: +420 377632510, e-mail: gruber@kky.zcu.cz

[2]Ing. Milan Legát, University of West Bohemia in Pilsen, Faculty of Applied Sciences, Department of Cybernetics, Univerzitní 22, 306 14 Pilsen, tel.: +420 377632510, e-mail: legatm@kky.zcu.cz

[3]Ing. Daniel Tihelka, Ph.D., University of West Bohemia in Pilsen, Faculty of Applied Sciences, Department of Cybernetics, Univerzitní 22, 306 14 Pilsen, tel.: +420 377632531, e-mail: dtihelka@kky.zcu.cz

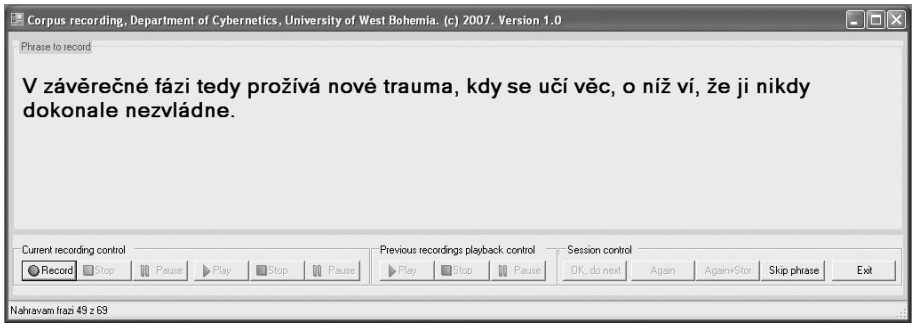

Fig. 1: The window of recording program.

2.2 Checking modules

Apart from the basic interface, several modules were created in order to control the features of recorded signals.

- Module for pause length checking.
- Module for intensity checking - used for both speech signal and glottal signal.
- Module for laryngograph error checking (battery discharge).
- Module for low frequency noise checking.

2.2.1 Pause checking

In the unit inventory there also have to be the units which represent pauses at the beginning and at the end of a spoken speech. For this purpose, the speaker has to keep pauses at the edges of utterances. Since there is a need to have some minimal duration of these units, a modul checking on the length of pauses was created. One second was exacted from the speaker as the minimum length of the pauses at both the beginning and the end of the utterance.

The pause check could be done in two basic ways. On the basis of either RMS[4] value of the signal or the maximum value of the signal. These two approaches can be combined, so we have one more possibility for this check.

First, we cut off some part from the beginning and from the end of the recorded speech, in order to separate the keyboard noise, which occurs at the edges of the utterance (this noise is shown in fig. 2). Then we check whether during the first and last second of the utterance the values of the signal do not exceed the fixed range. This is done by one of three possibilities mentioned above.

If the recorded utterance pass this procedure, the next checking module is employed, otherwise the speaker is forced to record the sentence again, until the result is correct.

[4]RMS = Root Mean Square, also known as the quadratic mean; a statistical measure of the magnitude of a varying quantity. It is especially useful when variates are positive and negative, e.g. waves.

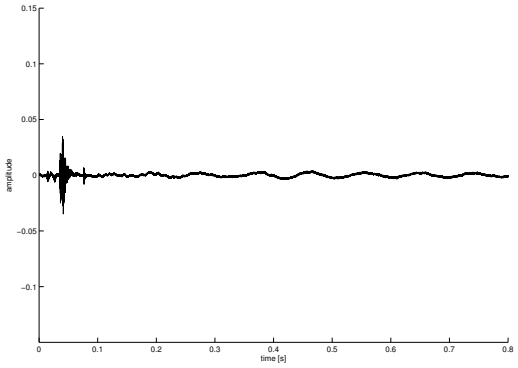

Fig. 2: A keyboard noise at the beginning of each recorded utterance. It occurs also at the end of it.

2.2.2 Intensity checking

When concatenating units from the unit inventory, as smooth joint as possible is required. If the neighbouring units have different amplitude level, an audible discontinuity visible in the waveform and also in the spectrogram could appear. In fig. 3 an example of this kind of concatenation is shown.

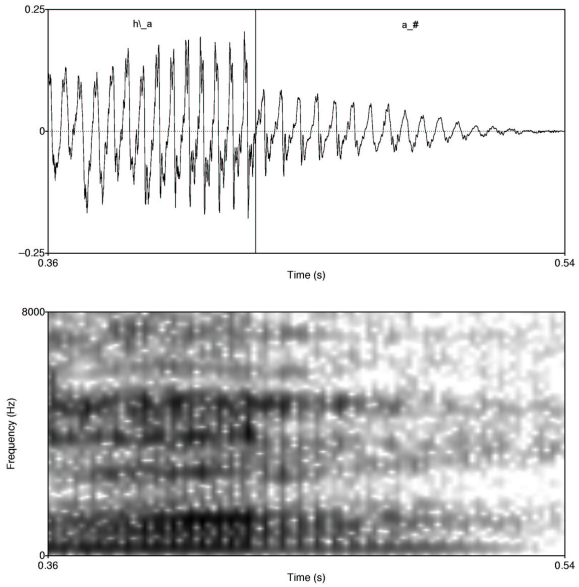

Fig. 3: Concatenation of two units (diphones) with different amplitude level.

In order to avoid this problem (at least partially), we demand the intensity of the speech and glottal signal to be similar in all sentences. Therefore another two modules checking on the signal intensity were created (one for speech channel, one for glottal channel).

Moreover, this check also ensures that values of the signal fall into given range. The recordings are required to be neither too silent nor too loud.

In order to perform this check we can again use three possibilites. The maximum value checking, the RMS value checking or combination of these two approaches. We

have to cut off the initial and final pause from the recorded utterance and compute the desired value for the rest of the signal. Then, this value is compared with fixed range and a conclusion is drawn. If the computed value fall into this range, the utterance passes the check and other one is performed, otherwise the speaker is informed that the utterance is too loud or too silent (the channel which is wrong is specified) and he is forced to record it again.

In some cases, the speaker is not able to control the intensity of the signal, especially when an error occurs in the glottal signal. The supervisor's role is then to adapt the settings of the recording equipment for particular utterance in order to meet the requirement for the intensity level.

2.2.3 Glottal signal checking

After recording previous speech corpus we found out, that the glottal channel in some sentences is damaged. This was caused by battery discharge in laryngograph, which is used for measuring the glottal signal. In order to avoid this problem in the new corpus, we decided to apply an automatic check on the glottal channel so that the battery discharge is revealed, the speaker and the supervisor are warned and further recording is not allowed.

The typical damaged segment of the glottal waveform is depicted in the upper part of fig. 4. Considering this kind of signal deterioration, we have proposed very simple detection method based on difference function and thresholding. The undamaged glottal signal is normally smooth and can be characterized by slightly ascending and sharply descending edges. The descending edges correspond with glottal closures and can be effectively used for pitch marking, as was shown in Legát (2007). On the other hand, in the damaged glottal waveform additive noise can be observed. This noise is not limited to be low or high frequency like and occurs randomly along the whole utterance.

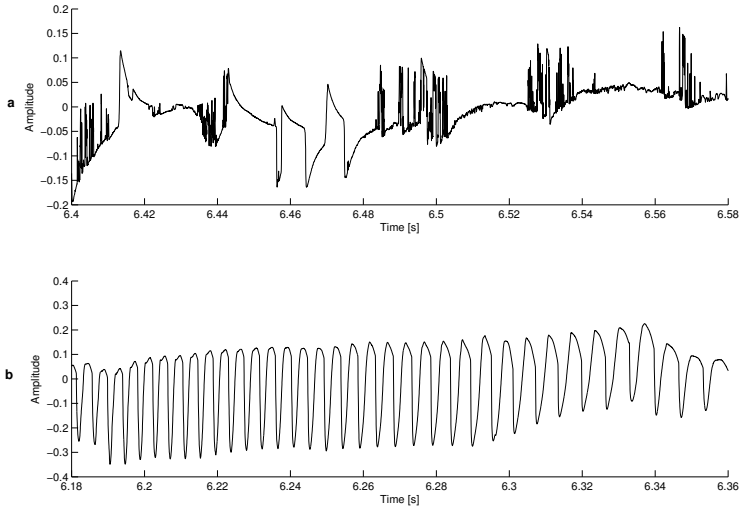

Fig. 4: a) a part of damaged glottal signal, b) a part of correct glottal signal

Our method for detection of segments which contain this kind of noise works in several steps. First we calculate the difference of the input glottal signal, let's denote the obtained signal *diffSignal*. As we search for abrupt ascending edges which are not normally present in glottal waveform, we perform the thresholding of the *diffSignal*. All the values lower than *diffThresh* are set to zero. The next step is to calculate the RMS value of the glottal signal frame by frame, the length of each frame is $10ms$ ad hoc. Then, we multiply the

values of thresholded *diffSignal* by the inverse of corresponding RMS frame values. By this multiplication we point the segments of the utterance where some sharply ascending edges are present, let's denote this vector *rmsDiffSignal*. To avoid the problem of marking undamaged segments as corrupted we set another threshold, *weightedDiffThresh*, and we set all the values of *rmsDiffSignal* below this threshold to zero, resulting in *weighted-DiffSignal*. The warning message about the glottal signal damage is printed if the sum of the values of the *weightedDiffSignal* is significantly higher than zero. All the constants used in this simple method need to be tuned on the basis of the intensity of the glottal signal. In our case, the values of glottal signal fall in the range $\langle -1, 1 \rangle$ and the following values of thresholds were used *diffThresh*=0.04 and *weightedDiffThresh*=0.6.

If the glottal signal passes this check, it is marked as correct. However, one more check is performed on the speech channel.

2.2.4 Low frequency noise checking

In a part of the previous corpus, some low frequency noise was detected (shown in fig. 5), which was probably caused by an improper setting of a microphone. For elimination of this deterioration another check module was created. It should reveal this kind of error in the speech signal.

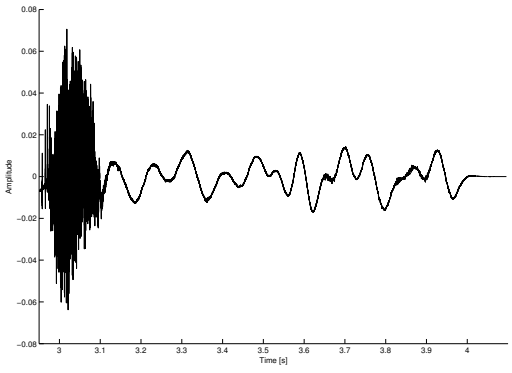

Fig. 5: Low frequency noise is especially visible at the end (or at the beginning) of an utterance where a silence should normally be. The end of speech and the noise is shown here.

This checking is provided by an analysis of the signal. At the beginning of the signal, where a pause should be, a zero-cross value is computed. This value differs in situations when there is an error or there is not. An average value in errorless sentences was determined and this value (with some range) was used as reference, which is compared with the computed one.

If the recorded utterance passes also this check, it is marked as the correct one and the next sentence is allowed to be recorded.

3 THE RECORDING

The recording was carried out in a consistent natural style by male speaker with some radio-broadcasting experience. The database of declarative and interrogative sentences contains 18 hours of natural speech stored in 11762 utterances. The whole recording was accomplished within three weeks.

The greatest problem regarding the checking modules was to find the optimal setting of their parameters beforehand. It is needed to tune all the parameters of the modules (the fixed values, the ranges, the thresholds, etc.) for the particular speaker. Anyway, the problem of the intensity checking of the glottal signal was described in 2.2.2.

Every recording day, there were three phases of the recording.

- Listening phase. The speaker was listening to the previously recorded sentences in order to remind his speaking style. Several sentences were used for this phase.
- Style adjusting phase. During this phase, speaker had to firstly record a given utterance (which was chosen from utterances previously already recorded) and after that he had to compare these two utterances (one recorded previously, one recorded anew) in terms of the speaking style. If he found the speaking style similar, he could go on to the recording phase. Again, several sentences were used for this phase.
- Recording phase. In this phase, the recording as such was performed.

4 CONCLUSION

During this recording we tried to check on the quality of the signals automatically (the glottal one and the speech one), so that we could reveal some errors that occured previously. For the further recording we need to tune the parameters of the modules more precisely in order to completely avoid the supervisor's intervent into the recording process. Some methods of modules could also be modified to obtain better results.

Acknowledgement: The work has been supported by GAČR 102/06/P205, by the EU 6th Framework Programme IST-034434 and by the Ministry of Education of the Czech Republic, project No. 2C06020.

REFERENCES

Matoušek, J., Romportl, J., 2007. *Recording and Annotation of Speech Corpus for Czech Unit Selection Speech Synthesis*, In: Matoušek, V., Mautner, P. (eds.) TSD 2007. LNAI 4629, pp. 326-333, Springer-Verlag Berlin Heidelberg.

Legát, M., Matoušek, J., and Tihelka, D., 2007.*A Robust Multi-Phase Pitch-Mark Detection Algorithm*, Proc. INTERSPEECH. Antwerp, Belgium, pp. 1641-1644