THE ISSUE OF CHECKING THE VOLUME CONSISTENCE OF SPEECH CORPUS DURING RECORDING

Milan LEGÁT[1], **Martin GRŮBER**[2], **Jindřich MATOUŠEK**[3]

Abstract: This paper deals with the problem of checking the consistency of the speech corpus during recording in terms of the level of speech power of individual recordings. The question was whether or not setting of the limits of RMS value is useful for checking the volume consistency of recordings destinated for unit selection speech synthesis.

Keywords: speech synthesis, unit selection, corpus recording, speech power

1 INTRODUCTION

Without question, the concatenative speech synthesizers can produce very well sounding and intelligible synthetic speech. These systems use the database of pre-recorded sentences from a single speaker, so called speech corpus. In speech generation process the appropriate units (i.e. phones, diphones, triphones etc.) are chosen from the database and concatenated resulting in synthetic utterance.

The approaches of concatenative speech synthesis can be divided into two groups – using single instance per each unit and using multiple instances of each unit, so called unit selection. Unit selection has become popular recently, because it can produce synthetic speech with very high naturalness, in comparison with single instance synthesis. The naturalness comes from the fact that the chosen units are concatenated directly and no or only slight signal processing is applied. Unfortunately, one of the features of unit selection is that now and then some unexpected glitches can occur in the synthetic utterance. There is a close relation between the quality of synthetic speech and the quality of speech corpus used for synthesis. Besides the typical factors such as segmentation accuracy, the number of instances of each unit or prosodic richness, the consistency of the speech corpus is crucial for the quality of synthetic speech.

In literature, there can be found some approaches to monitor the quality of speech corpus during recording. For instance, in Reller (2005) the tool for monitoring phone and diphone quality was presented. The quality of units was measured using spectral features. Another approach is to employ an expert in acoustic phonetics and orthoepy to supervise the recording of the speech corpus, Matoušek and Romportl (2007). However, it proved true that the expert is not able to guarantee the consistency of recordings in terms of speaker's style and the constancy of speaker's voice quality.

We have organized this paper into four sections. In section 2, we briefly outline the problem of speech corpus volume consistence checking during recording. Section 3 describes our experiments based on correlation analysis and it also serves to discuss the results. In section 4 we draw some conclusions and suggest some plans for future work.

2 VOLUME MONITORING TOOL

As mentioned above, the human expert is not able to guarantee the consistency of the speech corpus. Because of this fact, we have implemented a tool for monitoring the

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

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

[3] Ing. Jindřich Matoušek, Ph.D., University of West Bohemia in Pilsen, Faculty of Applied Sciences, Department of Cybernetics, Univerzitní 22, 306 14 Pilsen, e-mail: jmatouse@kky.zcu.cz (supervisor)

quality of recordings. This monitoring tool consists of several checking modules and one of these modules deals with the problem of volume consistency of the recordings. Before the implementation of this module, we discussed the question of how should the loudness of recordings be monitored or checked. Finally, we have implemented two criteria – peak based and RMS based. However, during the recording sessions, we experienced that the speaker had problems to meet the volume criterion in case of shorter sentences, especially questions. This phenomenon lead to the idea of performing the correlation analysis on recorded data (see sec. 3.1).

The idea of both volume criteria is very simple. For peak based criterion, we set the range into which the global maximum of the recorded speech waveform needs to fall. RMS based criterion works similarly, the only difference is that the claim is posed on the RMS value of the recorded sentence. Both of these approaches have some drawbacks. It is quite obvious that the peak based criterion tends to be too much local and does not guarantee the constancy of speech power along the sentence. For instance, this criterion can be satisfied even if only the first word of the sentence is loud enough and the rest of the sentence is rather silent. This is the thing we need to avoid when the corpus is designed for the unit selection.

RMS based criterion is more robust regarding this phenomenon, but still it is not able to meet all the requirements we pose on the quality of the speech corpus. The problem is that the RMS based criterion seems to be too much global, compared with peak based one. This can be explained by small example. Let us imagine the situation we have two sentences – short one (e. g. consisting of three words) and long one, which can be in fact compound sentence consisting of more sentences. It is natural that the long sentence contains some pauses as the speaker needs to have some breaks for breath. If we calculate the RMS value of this compound sentence from the beginning to the end of the utterance these pauses are included. The result is that the calculated RMS value of the whole sentence is lower than the real RMS value of the wanted signal in which we are interested. It means that if we set the range of RMS values into which we want all the recorded sentences to fall, we force the speaker to read the longer sentences more loudly than the short ones which do not contain any pauses. Hence, with regards to the segmental level the short and long sentences are not equal in terms of segmental loudness. In fig. 1 there is shown the concatenation of two units of different loudness. This kind of concatenation is perceived by the listeners as glitches.

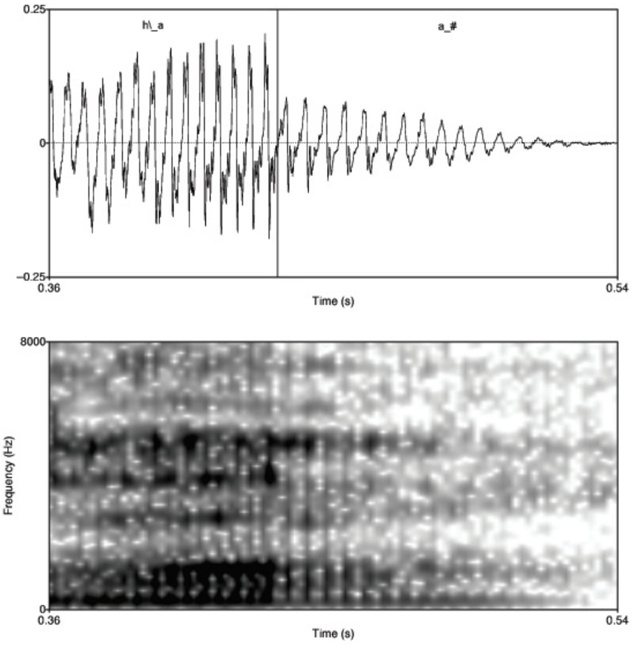

Fig. 1: Concatenation of units with different level of loudness - waveform and spectrogram.

3 CORRELATION ANALYSIS

In the previous paragraph, we have purposely used the word seem when we were discussing the problem of RMS based criterion. Actually, at the time of developing of the volume checking module, we did not know whether or not the presence of pauses in long sentences influences the RMS value of the whole sentence significantly. Moreover, there was one more question relating to the ratio of strong sounds (i. e. vowels and sonorous sounds) in the sentence and its impact on the overall RMS value of the utterance.

3.1 Correlations to analyze

To testify or disprove our assumptions by evidence, we have used two speech corpuses (Sec. 3.3.1) and performed correlation analysis on them. We have tested following correlations:

1. The correlation between ratio value of strong sounds and RMS value of the sentence. The ratio value of strong sounds was defined, as follows:

$$ratio\ value = \frac{number\ of\ vowels\ +\ number\ of\ sonorous\ sounds}{number\ of\ all\ sounds} \tag{1}$$

2. The correlation between the length of the sentence and RMS value of the sentence.

3.2 Correlation analysis in brief

The word “correlation” represents the rate of association of two variables. We can say that two variables are correlated when the certain values of the first one tend to concur with certain values of the second one. Two variables are absolutely correlated when the certain value of the variable X concur with exactly one certain value of the variable Y. Contrariwise, if the probabilities of all the values of the variable Y concurring with the certain value of the variable X are equal, we say that these to variables are noncorrelated. There are many coefficients for measuring the dependence between random variables. The most frequently used one is Pearson's correlation coefficient r, even if it has some drawbacks. This coefficient indicates the strength and direction of a linear relationship between two random variables. The values of the coefficient fall into the range $\langle -1, 1\rangle$. The closer the coefficient is to either -1 or 1, the stronger the correlation between the variables is. If it is equal to 1 (or -1) the exact linear relation between the given variables can be found.

Strength of correlation	r
weak	0.1-0.3
medium	0.3-0.7
strong	0.7-1.0

Tab. 1: The strength of association of variables according to the correlation coefficient r.

Having series of n measurements of X and Y written as x_i and y_i, where $i = 1, \ldots, n$, the Pearson's correlation coefficient r_{xy} is written:

$$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{n-1} \tag{2}$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \tag{3}$$

where s_{xy} is the covariance of the variables X and Y, s_x and s_y are the sample standard deviations. If the value of r_{xy} is positive the association between given variables is also positive and vice versa.

3.3 Experiments

3.3.1 Data used in analysis

To analyze the correlations mentioned above (see sec. 3.1), we have used two recorded speech corpuses. One comes from the female speaker and one from male speaker. Recording of the female speaker was supervised by human expert and no automatic monitoring tool was used to monitor recording sessions. Contrariwise, recording of the male speaker was unsupervised and the automatic monitoring tool was employed instead of the human expert. The loudness of these recordings was checked by the volume module and both peak based and RMS based criterion were used. In addition, we have recorded 341 interrogative sentences using only the peak based volume criterion.

3.3.2 Experimental results

In tab. 2 there is the summary of the results of the correlation analysis performed on the data recorded by the male speaker. The abbreviation $quest_{rms}$ stands for interrogative questions which were checked using RMS based criterion, *quest* stands for interrogative questions checked by peak based criterion and $decl_{rms}$ are the declarative sentences monitored by RMS based criterion. The rows of the table denoted *ratioValue* and *uttLength* represents the correlations of our interest.

	quest$_{rms}$	quest	decl$_{rms}$
ratioValue	0.2103	0.3190	0.0308
uttLength	-0.2298	-0.2089	-0.0886

Tab. 2: Correlations calculated for male speaker.

We have calculated mean values and standard deviations of the variables involved in the correlation analysis (see tab. 3) to find out whether or not the obtained correlations are comparable. Having a look at this table, we can see that declarative sentences involved in the analysis were cosiderably longer than interrogative sentences.

	quest$_{rms}$		quest		decl$_{rms}$	
	mean	std	mean	std	mean	std
ratioValue	0.4599	0.0597	0.4573	0.0632	0.4675	0.0430
uttLength	2.5086	1.6354	2.4018	1.578	4.5549	1.7928
rmsValue	0.17	0.0223	0.1609	0.0168	0.1474	0.0150

Tab. 3: Mean values and standard deviations of observed variables – male speaker.

To obtain comparable sets of sentences in terms of their length we have selected a subset of longer interrogative sentences (denoted *longQuest*) from both sets of interrogative sentences. The results of correlation analysis performed on these sets are shown in tab. 4. We have also separated some shorter interrogative sentences and the summary of results can be found in the same table. The mean values and standard deviations of variables measured on these subsets are summarized in tab. 5.

As mentioned above, we have performed the same experiments on another speech corpus, this one contains sentences from female speaker. The results of the correlation analysis are summarized in tab. 6 and tab. 7.

3.3.3 Discussion of experimental results

Having a look at RMS values of recorded male sentences $quest_{rms}$ and $decl_{rms}$, we were supprised that these are not equal. The first on hand explanation of this fact was this difference in overall RMS values is caused by presence of pauses in longer sentences. But this is obviously not true, because the same thing can be observed when we have a look at overall RMS values of longer interrogative sentences (*longQuest*). Moreover, the analysis of loudness of vowels in all sets (see tab. 8) shows that interrogative sentences are really

	shortQuest$_{rms}$	shortQuest	longQuest$_{rms}$	longQuest
ratio value	0.2190	0.3443	0.2044	0.3449
utt length	-0.2582	-0.0415	-0.1149	-0.1709

Tab. 4: Correlations calculated for male speaker - long ($\geq$ 3s) and short ($\leq$ 2s) questions.

	shortQuest$_{rms}$		shortQuest		longQuest$_{rms}$		longQuest$_{rms}$	
	mean	std	mean	std	mean	std	mean	std
ratioValue	0.4609	0.0720	0.4640	0.0767	0.4618	0.0373	0.4517	0.0419
uttLength	1.3478	0.3689	1.2947	0.3728	4.7106	1.5005	4.5294	1.4757
rmsValue	0.1746	0.0245	0.1641	0.0171	0.1639	0.0180	0.1567	0.0164

Tab. 5: Mean values and standard deviations of observed variables in long ($\geq$ 3s) and short ($\leq$ 2s) questions – male speaker.

	quest	decl
ratioValue	0.1684	0.145
uttLength	-0.1994	-0.1416

Tab. 6: Correlations calculated for female speaker.

	quest		decl	
	mean	std	mean	std
ratioValue	0.474	0.061	0.4734	0.0494
uttLength	4.1989	3.7384	4.8302	3.2714
rmsValue	0.1497	0.0243	0.1493	0.017

Tab. 7: Mean values and standard deviations of the observed variables – female speaker.

a bit louder, even though the set *quest* was not checked by RMS value criterion during recording. It seems that the only explanation of this phenomenon is that the speaker changed his style when recording interrogative sentences and the volume criterion was not as strict as it should be to guarantee the consistency of speech power along all sessions.

	quest$_{rms}$		quest		decl$_{rms}$	
	mean	std	mean	std	mean	std
a	0.2282	0.0471	0.2129	0.0430	0.1955	0.0366
e	0.2245	0.0367	0.2088	0.0345	0.1964	0.0366
i	0.1921	0.0409	0.1787	0.0398	0.1640	0.0319
o	0.2224	0.0437	0.2111	0.0396	0.1957	0.0364
u	0.1755	0.0553	0.1586	0.0439	0.1488	0.0375

Tab. 8: Comparison of loudness of recorded vowels – male speaker.

Regarding interrogative sentences, another interesting thing was that there was also difference in loudness between sets *quest*$_{rms}$ and *quest*. To find out what the reason of this fact is, we analyzed sentences which did not meet the RMS volume criterion during recording. By this slight analysis we discovered the mistake in setting of the volume checking module. The problem was that the module tended to mark starts and ends of utterances erroneously and short segments of silence were added to the overall RMS value. Because of this fault in module setting the speaker was forced to speak louder when recording shorter sentences, i.e. probably majority of interrogative sentences checked by

RMS volume criterion. This is also the explanation for the fact that the speaker had problems to meet the volume criterion during recording of shorter sentences.

Having in mind this problem of module setting, we can move on to the results of the correlation analysis. Some useable results can be obtained from the set of interrogative sentences uttered by male speaker (*quest*) as this one was not checked by RMS value criterion during recording. By the analysis performed on this set we have validated our preliminary idea that there is a correlation between ratio of strong sounds (vowels and sonorous sounds) in sentence and the overall RMS value of the utterance. This correlation was partly damaged in the set of interrogative sentences checked by RMS value criterion ($quest_{rms}$). The explanation is that as the speaker was forced to speak louder (because of the problem with checking module setting) he was probably reaching the upper RMS value limit when recording sentences with higher ratio value and was forced to change his style to meet the volume criterion. There is no correlation in the set of declarative sentences ($decl_{rms}$). The cause of this fact can be the presence of both short and long sentences in this set. However, the further analysis would be required to confirm this presumptions. In case of female speaker there is also weak correlation between these variables, although it is not so obvious, compared with male speaker.

As regards the correlation between length of the utterance and its overall RMS value, the idea of presence of weak negative correlation between these variables was partly confirmed. The surprising result was obtained for the set of declarative sentences uttered by male speaker. Unfortunately, we are not able to explain why there is no correlation in this set.

4 CONCLUSION

In this paper we have addressed the question of how the volume consistency of the speech corpus destinated for unit selection speech synthesis could be checked during recording sessions. We have experienced the change of speaking style of male speaker when recording interrogative sentences. This observation comes from comparison of vowel intensities in declarative and interrogative sentences which were not monitored by RMS value based criterion. From this point of view, it seems suitable to set limits for overall RMS value of utterances.

On the other hand, there is a weak correlation between the ratio of strong sounds (vowels and sonorous sounds) in sentence and the overall RMS value in interrogative sentences uttered by male speaker. Hence, setting of the range for overall RMS values of utterances need to be the trade-off and the further analysis would be required to answer this question satisfactorily. According to our experiments, the female speaker was more consistent during recording sessions in terms of speech power in interrogative and declarative sentences.

Future work will focus on the question how the differences in loudness can influence the perception of synthetic speech and incorporation of the results into our speech synthesis system. More sophisticated approach of checking the volume consistence of speech corpus during recording will be also researched.

Acknowledgement: This research was supported by MSMT, project No. 2C06020, and the EU 6th Framework Programme IST-034434.

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.

Reller, Ch., 2005. *Diphone Corpus Recording and Monitoring Tool*, Thesis, Institut für Technische Informatik und Kommunikationsnetze, Zürich.