REDUCING FOOTPRINT OF UNIT SELECTION TTS SYSTEM BY EXCLUDING UTTERANCES FROM SOURCE SPEECH CORPUS

Jindřich Matoušek, Daniel Tihelka, Zdeněk Hanzlíček

Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia
Univerzitní 8, 306 14 Plzeň, Czech Republic
jmatouse@kky.zcu.cz, dtihelka@ kky.zcu.cz, zhanzlic@ kky.zcu.cz

Abstract
Current unit selection speech synthesis systems are capable of producing speech of a high quality at the expense of enormous computational and storage requirements. In this paper, the analysis of an existing large speech corpus employed for unit-selection-based synthesis of Czech speech is performed. Subsequently, a procedure for the exclusion of some amount of utterances from the source speech corpus is proposed. The procedure is based on the statistics of the utilisation of all utterances during text-to-speech synthesis of a large portion of texts. The exclusion of whole utterances was preferred over the exclusion of the particular instances of speech units in order to preserve the main feature of unit selection framework – to select as longest sequence of contiguous speech units as possible. After the exclusion, the footprint of the system was reduced approximately by 42 %. The resulting synthetic speech was then judged by means of 5-scale CCR listening tests and evaluated in average as only "slightly worse" than speech generated by the baseline (i.e. not reduced) system.

1 Introduction

The current trend in speech synthesis is to use large carefully prepared speech corpora comprising many instances of each speech unit and unit selection techniques to select the optimal sequence of unit instances when producing the output speech. As the resulting speech is made up by concatenating pre-recorded segments of natural speech, such approaches (also known as corpus-based speech synthesis) are able to generate speech of a high quality [1]. On the other hand, the computational requirements are enormous (including the storage issues, so called footprint – hundreds of megabytes of RAM are usually required), preventing the technology to be utilised on less powerful or low-resource devices like pocket PCs, mobile phones, etc., or even on server-like systems where more voices (i.e. more corpora) are to be stored. There are two main issues which should be addressed when analyzing the system demands – the storage (memory) requirements (*footprint*) and the runtime *computational* requirements which correspond to the speed of selecting the appropriate instances of speech units from the speech corpus. Although both issues are very important, the former one will be further researched in this paper. More specifically, we will focus on the reduction of the footprint of the system by excluding some amount of the utterances from the source speech corpus.

The paper is organised as follows. The Section 2 introduces the baseline text-to-speech (TTS) system. In Section 3 the experiments with the reduction of the system footprint are described, including the procedure for the exclusion of the utterances from the source speech corpus. Section 4 then presents the evaluation of the reduced system and Section 5 concludes the paper.

2 The Baseline System

In our experiments we employed the Czech text-to-speech (TTS) system ARTIC (Artificial Talker in Czech). More specifically, the unit selection module of ARTIC basically as described in [2] was utilised. Based on a carefully designed speech corpus (annotated on orthographic, phonetic and prosodic levels [3]), statistical approach (employing hidden Markov models, HMMs) was employed to perform the automatic phonetic segmentation of the source speech corpus into phones [4]. Based on this segmentation, boundaries between diphones, the basic speech units used in the ARTIC unit selection system, were located. As a result, *acoustic unit inventory* (AUI), the source speech corpus indexed with diphones and prosodic structures [5], was built.

During the runtime synthesis the phonetic and prosodic aspects of the input text are estimated and, based on these features and also on the acoustic contexts of the surrounding units, the optimal sequence of each diphone instances is selected. Output speech is then made up by concatenating the selected diphone instances. The prosodic characteristics of the synthesised speech are controlled at the symbolic level, utilising general linguistic features like the position of each diphone in different prosodic structures (e.g. word, phrase, clause or the whole utterance), the type of prosodeme the diphone is in [5], etc. The block diagram of the ARTIC TTS system is shown in Figure 1.

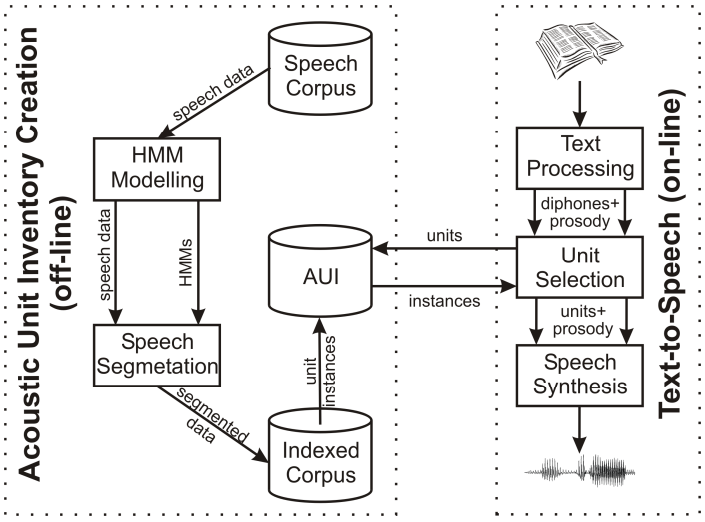

Figure 1 The simplified scheme of the baseline text-to-speech system.

The speech corpus used in our experiments comprises 12,242 utterances of a male voice (almost 15 hours of speech excluding utterance leading and trailing silences). The utterances were carefully selected in order to be both phonetically and prosodically balanced [3]. The types of the utterances and their numbers are shown in Table 1. The resulting number of all diphone tokens (i.e. the total number of all instances of all diphones) is 670,757. The speech waveforms representing all diphone instances were coded using 5-bit ADPCM coding scheme. The total footprint of the original system was 670 MB.

Type of utterances	**Number**
declarative	10,009
yes/no questions	1,002
“wh”-questions	700
application-based	531

Table 1 The overview of the types and numbers of utterances in the speech corpus.

3 Experiments

There are several methods dealing with the reduction of the footprint of a TTS system described in literature. They mostly concern various speech coding techniques and aim at reducing the footprint by employing one of the coding techniques, e.g. [6]–[8]. From the point of view of the preservation of the speech unit instances in the acoustic unit inventory, such approaches are lossless – no reduction of the number of unit instances is performed at all, the footprint is reduced by utilising a speech coding method.

In our experiments, besides the ADPCM speech coding technique, the system footprint is reduced by excluding some speech unit instances from the source speech corpus. Moreover, the possibilities of excluding *whole utterances* (and all speech unit instances included in them) are researched in this work. The exclusion of whole utterances was preferred over the exclusion of the particular instances of speech units in order to preserve the main feature of the unit selection framework – to select as longest sequence of contiguous speech units as possible. Such an approach perfectly fits in the concept of the ARTIC TTS system, in which minimum, or even no modifications of the synthesised speech signal are carried out. However, it should be noted that such an approach is lossy, because some speech unit instances are thrown away.

It is obvious that synthetic speech produced by either approach can suffer from the reduction. Most speech coding techniques work in frequency domain and the employment of a speech coding model often results in a "buzzy-like" sounding speech. In Section 4 we will evaluate the impacts of the footprint reduction by excluding the whole utterances on the quality of synthetic speech.

3.1 The Algorithm

The main idea of the proposed algorithm is quite naive. It can be assumed that the fewer utterances are present in the speech corpus, the smaller is the size of the acoustic unit inventory. Usually, it is hard to experiment with different speech corpora or with different parameters of a single corpus (e.g. various kinds of phonetic/prosodic distribution). The process of obtaining a new speech corpus is very lengthy and expensive, so recording several corpora of the same voice is not effective. Hence, as much utterances as possible are usually recorded given speaker's capabilities and funding available.

The task of the proposed algorithm is to select utterances which will be excluded from the source speech corpus. The utterances to exclude were selected mainly on the basis of the statistics of their utilisation in speech synthesis of a large portion of texts. Such an approach was preferred over the random selection in which some of the very often employed utterances could be possibly omitted. The algorithm works in three phases. The details about each phase are given in the next sections.

3.2 Phase 1: The Exclusion of "Special" Utterances

In the 1st phase, all application-based utterances (utterances from very special domains such as railway stations, various call centres, etc.) and "wh"-questions (both denoted here as "special" utterances) were excluded. Such utterances are considered not so important in a general-purpose TTS system. In addition, "wh"-questions (unlike yes/no questions) are prosodically similar to declarative utterances. Therefore, "wh"-questions could be to some extent synthesised from declarative utterances. After the exclusion of these utterances, the number of utterances was reduced to 11,011.

3.3 Phase 2: The Exclusion of Utterances with Poor Segmentation Results

As it is well-known that speech units with poorly segmented boundaries degrade the quality of resulting speech when used during unit selection speech synthesis, the automatic phonetic segmentations of the rest of the utterances were analysed and utterances with potentially poor segmentation results were excluded in this phase.

Three criteria were employed for the detection of the utterances with poor segmentation results:

- the presence of a phone with a very long duration (more than 400 ms);
- the presence of a phone with a very short duration (less than 12 ms);
- the presence of a phone with a very low segmentation score (less than -120 log probability, see [9] for the description of the segmentation score).

Utterances with at least one phone segment with "suspiciously segmented" boundaries as indicated at least by one of the three criteria were excluded. After the exclusion, 8,260 declarative utterances and 862 yes/no questions remained in the corpus.

3.4 Phase 3: The Exclusion of the Rarely Employed Utterances

In the last phase, the analysis of the utilisation of the rest utterances was carried out. Approximately 524k text sentences were synthesised by the baseline TTS system and a record of how many times each utterance from the source speech corpus was employed during the synthesis was stored for further analyses. The analysis was carried out separately for declarative sentences and for yes/no questions. Since declarative utterances are employed more often than yes/no questions, the results of the joint analysis could be biased towards declarative utterances.

Having had the statistics of the usage of each source utterance during synthesis of the large portion of texts, the detection of outliers (the rarely used utterances in our case) was performed. However, the "standard" outlier detection techniques like the use of mean and standard deviation, p^{th} percentile or five-number summary would result only in a small reduction because only few utterances would be detected as "lower outliers" (818 utterances for the 10^{th} percentile statistics, no lower extremes were even detected for five-number summary).

To detect a "reasonable" number of "lower outliers", a coverage analysis was proposed. It consists of four steps:

1. Sort utterances according to their usage during synthesis of a large portion of texts from the most often used ones to the least often used ones. The usage is measured by means of the number of speech unit instances utilised during synthesis.
2. Compute the cumulative sum of the utilisation of each utterance.
3. Go through the cumulative sum and stop when the cumulative sum of an utterance reaches the required coverage. By the coverage we mean the percentage of the utilisation of the source utterances.
4. The utterances ensuring the given coverage are preserved. The rest of utterances are excluded, because they do not contribute to the speech synthesis by the original system so much.

The results of the coverage analysis for declarative utterances and for different coverage thresholds are shown in Table 2. For declarative utterances 90% coverage threshold was used and 2,187 utterances were excluded. The cumulative sum and the 90% coverage threshold are depicted in Figure 3. For yes/no questions 95% coverage threshold was employed and 304

Coverage	# Outliers
95 %	1,327
90 %	2,187
85 %	2,904
80 %	3,529

Table 2 The number or "lower outliers" for different coverage thresholds.

System	Original	Reduced
Utterances	12,242	6,631
Amount [h:m]	14:44	8:43
Size [MB]	670	389
Unit tokens	670,757	394,722

Table 3 The comparison of the original and the reduced systems.

utterances were excluded. At the end of this phase, the total number of 9,122 utterances was reduced to 6,631 utterances. The comparison of the characteristics of the original and the final reduced system is shown in Table 3. The footprint of the system was reduced by approx. 42%.

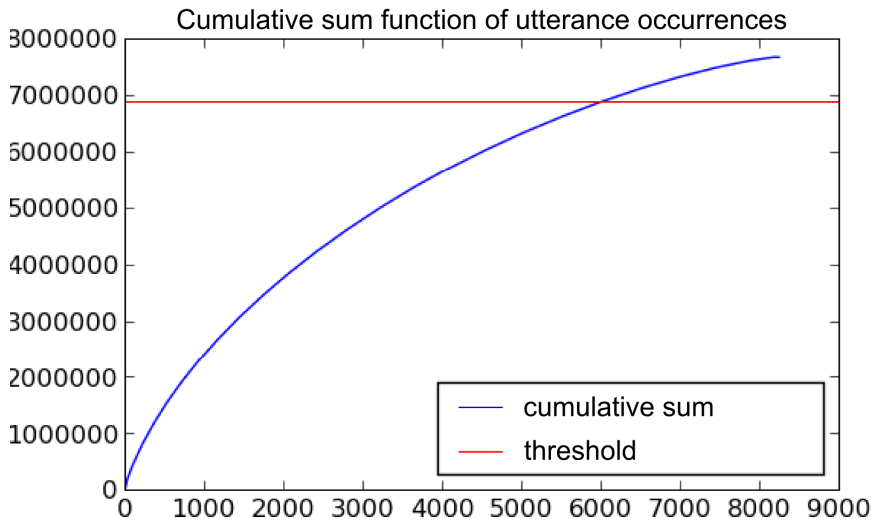

Figure 2 The cumulative sum and the 90% coverage threshold for declarative utterances.

4 Evaluation

The comparison of the quality of synthetic speech produced by both the original and the reduced TTS system was carried out by means of listening tests. As the test stimuli, the utterances affected by the reduction process the most were selected. All 524k text sentences were synthesised by both the original and the reduced system and the utterances which consisted of the most number of differences were selected for listening tests. Three criteria for the measurement of the differences were proposed:

- the number of different speech unit instances per utterance (DiffUnits);
- the increase of the concatenation points per utterance (IncConcat);
- the number of unit instances from the excluded utterances used for synthesis from the reduced TTS system (NumExclud).

All criteria were normalised by the length of the utterances. For each criterion, 10 utterances were chosen. So, there were 30 test utterances available for the listening tests in total. It should be noted that, because the most different utterances were selected, in fact, from the point of view of the reduced system, the worst possible cases were evaluated.

Five-point Comparison Category Rating (CCR) listening tests (specified in Table 4) were used for the evaluation. Five listeners experienced with speech synthesis participated in the tests.

Rate	Original system (O) compared to reduced system (R)	
2	O >> R	O much better than R
1	O > R	O slightly better than R
0	O = R	O equals R
-1	O < R	O slightly worse than R
-2	O << R	O much worse than R

Table 2 The specification of CCR listening tests used in the evaluation.

The results of the evaluation are shown in Figure 3. In average, speech synthesised by the reduced system was assessed as *slightly worse* (the average rate was 0.51). As can be seen, for the criterion DiffUnits the reduced system was evaluated as the same or even slightly better than the original system. On the other hand, for the other criteria the reduced system was evaluated as slightly worse. The results are statistically significant (p = 0.0001, sign test). The histogram of all individual assessments is shown in Figure 4.

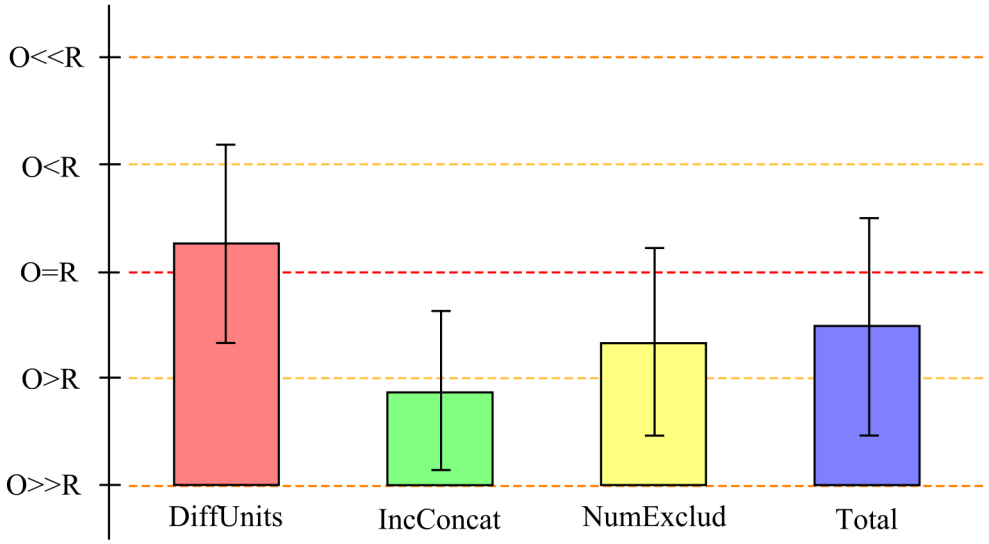

Figure 3 The comparison of the quality of synthetic speech produced by the original (O) and reduced (R) system by means of CCR listening tests for utterances selected according to three different criteria.

5 Conclusion

In this paper, the research into the reduction of the footprint of the Czech unit-selection TTS system ARTIC was described. The reduction was achieved by excluding specially selected utterances from the source speech corpus. The selection of the utterances to exclude was based on the coverage analysis (in sense of the utilisation of utterances from source speech corpus during TTS synthesis of very may text sentences). Utterances with poor automatic phonetic segmentation results and application-based utterances were also excluded. After the reduction, the number of utterances decreased from 12,242 to 6,631 which correspond to the reduction of the footprint of the TTS system by 42 % to 389 MB. The quality of synthetic

speech produced by the reduced system was evaluated by means of listening tests as slightly worse than the quality of synthetic speech produced by the original system in the worst possible cases. Although the experiments were carried out for the given coverage threshold (90 % for declarative utterances), they could be easily tuned to other coverage thresholds. In this way, a trade-off between the footprint of the system and the quality of resulting synthetic speech can be efficiently reached.

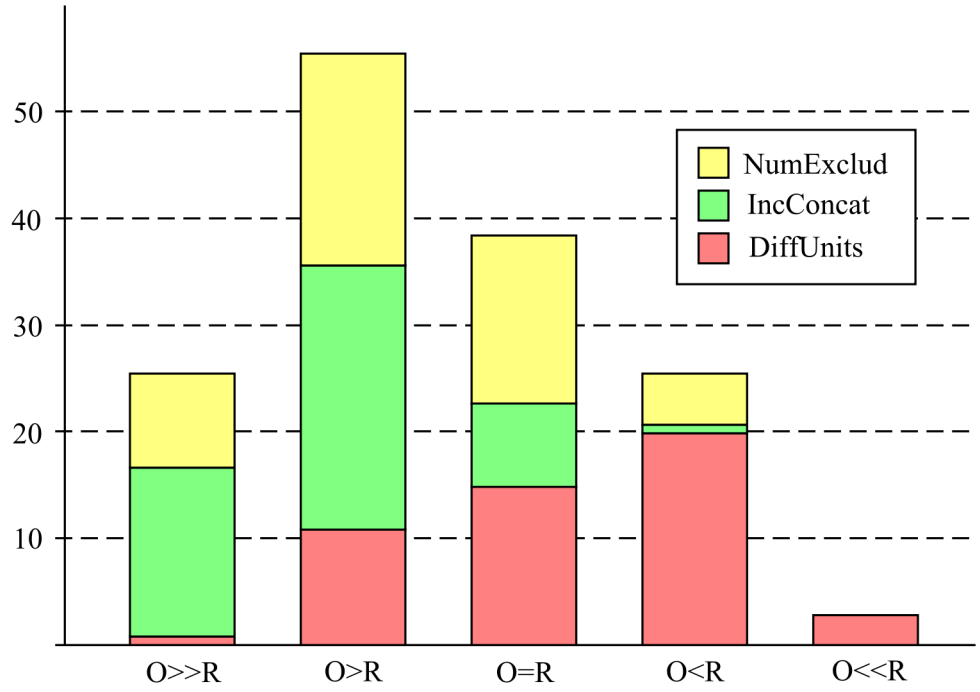

Figure 4 Histogram of all individual assessments of synthetic speech generated from utterances selected according to three different criteria.

6 Acknowledgements

Support for this work was provided by the Ministry of Education of the Czech Republic, project No. 2C06020, and by the Grant Agency of the Czech Republic, project No. GAČR 102/09/0989. The access to the METACentrum supercomputing facilities provided under the research intent MSM6383917201 is highly appreciated.

References

[1] Dutoit, T.: Corpus-based Speech Synthesis. In: Benesty, J., Sondhi, M., Huang, Y. (eds.) Springer Handbook of Speech Processing. Springer, Dordrecht 2008, pp. 437–455.

[2] Matoušek, J., Tihelka, D., Romportl, J.: Current State of Czech Text-to-Speech System ARTIC. Lecture Notes in Artificial Intelligence: Text, Speech and Dialogue, vol. 4188, Springer, Berlin 2006, pp. 439–446.

[3] Matoušek, J., Tihelka, D., Romportl, J.: Building of a Speech Corpus Optimised for Unit Selection TTS Synthesis. In: Proc. Int. Conf. on Lang. Resources and Evaluation (LREC), Marrakech, Morocco 2008.

[4] Matoušek, J., Romportl, J.: Automatic Pitch-Synchronous Phonetic Segmentation. In: Proc. Interspeech, Brisbane, Australia 2008, pp. 1626–1629.

[5] Romportl, J.: Statistical Evaluation of Prosodic Phrases in the Czech Language. In: Proc. Speech Prosody, Campinas, Brazil 2008, pp. 755–758.

[6] Chazan, D., Hoory, R., Kons, Z., Sagi, A., Shechtman, S., Sorin, A.: Small Footprint Concatenative Text-to-Speech Synthesis System using Complex Spectral Envelope Modeling. Proc. Interspeech, Lisbon, Portugal 2005, pp. 2569–2572.
[7] Lee, C.-H., Jung, S.-K., Eriksson, T., Jun, W.-S., Kang, H.-G.: An Efficient Segment-Based Speech Compression Technique for Hand-Held TTS Systems. Proc. Interspeech, Pittsburgh, USA 2006, pp. 213–216.
[8] Strecha, G., Eichner, M., Hoffmann, R.: Line Cepstral Quefrencies and Their Use for Acoustic Inventory Coding. Proc. Interspeech, Antwerp, Belgium 2007, pp. 2873–2876.
[9] Young, S., et al: The HTK Book (for HTK Version 3.4). Cambridge University, Cambridge, U.K. 2006.