Corpus Creation for Polish Unit Selection Speech Synthesis

1Krzysztof Szklanny, 2Dominika Oliver

1 Multimedia Department, Polish-Japanese Institute of Information Technology

Koszykowa 86, 02-008 Warsaw, Poland

2 Institute of Phonetics, Saarland University,

Building 17 Postfach 15 11 50

66041 Saarbrücken, Germany

kszklanny@pjwstk.edu.pl, dominika@coli.uni-sb.de

ABSTRACT

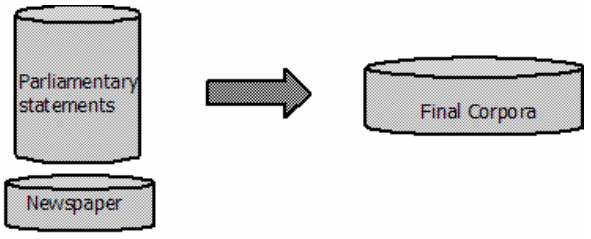
This paper describes the process of creating speech corpus for Polish Unit Selection speech synthesis. This task is time-consuming and manually designing the corpus is, in practice, only applicable in Limited Domain Speech Synthesis and Recognition. The sentence selection tools used while designing the corpus are usually based on the Greedy algorithm. The algorithm looks for sentences which cover the input parameters. The bigger the text set, the better the chance to fulfill given criteria. The main aim of this study is to design a speech corpus for Polish Unit Selection Speech Synthesis on the basis of phoneme, diphone and triphone frequency distribution. Research on using variable length units from different phonetic and prosodic contexts shows that when such units are joined together they help achieve natural sounding speech synthesis.

1. Introduction

To create the corpus, texts from parliamentary statements and newspaper reviews were used. First, existent corpora, see Picture 1, from these two domains had been used to carry out statistical analysis to determine the differences in their phonetic distribution. Despite the difference in domain and size ratio (10:1), neither of the data sources

differed significantly as far as the relative frequency of phonemes present were concerned.

Based on this analysis, the corpus of parliamentary statements was chosen as the initial corpus used for sentence selection. Another argument for using this data source is its size, namely, 300 MB which corresponds to 5778460 sentences. The necessary pre-processing of these sentences included removal of all the tags and other metadata. Next, abbreviations and numbers were expanded. Sentences in graphemic form had to be transformed into their phonetic transcription. In order to minimize processing time the corpus was divided into a dozen sub-corpora and then phonetic transcription was generated for each of them simultaneously. The phonetic transcription of phonemes diphones and triphones was derived using grapheme-to-phoneme conversion for Polish.

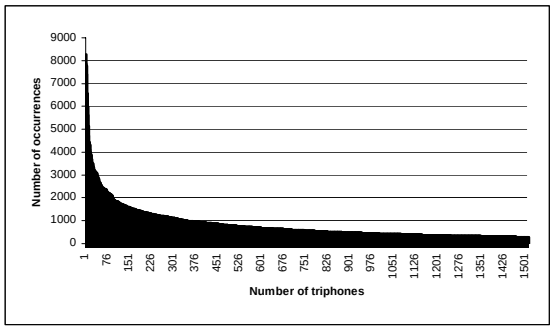


Picture 1. *Source data domain.*

2. Unit selection

The corpus selection is based on phonemes, diphones and triphones. Diphones and triphones are units which can be more easily and successfully joined than phonemes. The number of phonemes in Polish is 37 and there are 1443 diphones. Earlier experience with preparing a diphone database for Polish and using it in concatenative synthesis [7, 8] confirm that diphones help obtain natural sounding speech. As for triphones, they can also be easily concatenated but obtaining a full coverage for triphones is impractical because of the huge number of triphones [10].

Our analysis shows that there are 400 most frequent triphones in all sub-corpora used here. They occur at least 1000 times in the initial corpus. Picture 2 shows a distribution of 1500 most frequent triphones for all sub-corpora.



Picture 2. *Number of occurrences of most frequent triphones.*

An example input sentence in out initial corpus is in its orthographic and phonetic form represented by 1a) orthography 1b) phonemes 1c) diphones and 1d) triphones.

1a. jeśli chodzi o utrzymanie infrastruktury szacuje się potrzeby roczne

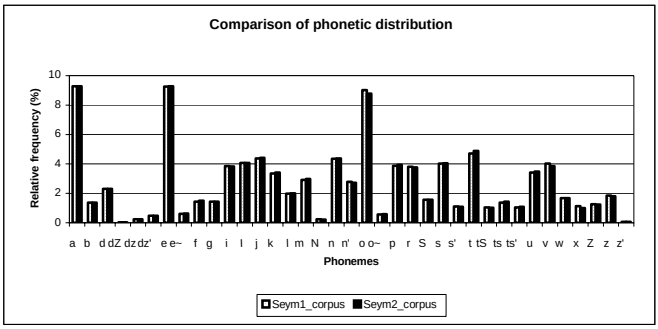
1b. #j e s' l i x o dz' i o u t S I m a n' e i n f r a s t r u k t u r I S a ts u j e s' e~ p o t S e b I r o tS n e#

1c. #j je es' s'l li ix xo odz' dz'i io ou ut tS SI Im ma an' n'e ei in nf fr ra as st tr ru uk kt tu ur rI IS Sa ats tsu uj je es' s'e~ e~p po ot tS Se eb bI Ir ro otS tSn ne e#

1d. #je jes' es'l s'li lix ixo xodz' odz'i dz'io iou out utS tSI SIm Ima man' an'e n'ei ein inf nfr fra ras ast str tru ruk ukt ktu tur urI rIS ISa Sats atsu tsuj uje jes' es'e~ s'e~p e~po pot otS tSe Seb ebI bIr Iro rotS otSn tSne ne#

3. Preparing the corpora

The initial corpus of parliamentary statements has been randomly divided into 12 sub-corpora. This was a necessary format requirement for the greedy algorithm program that was used. Each sub-corpus contained about 22000 sentences. Again the phonemes frequencies were similar for each of the sub-corpora. Picture 3 illustrates the comparison of phonetic distribution between two randomly selected sub-corpora.



Picture 3. *Comparison of phonetic distribution between two random sub-corpora of parliamentary statements.*

Next analysis involved deriving the same coverage statistics for the other domains of text corpora. They contained general newspaper texts as well as newspaper reviews (19733 sentences).

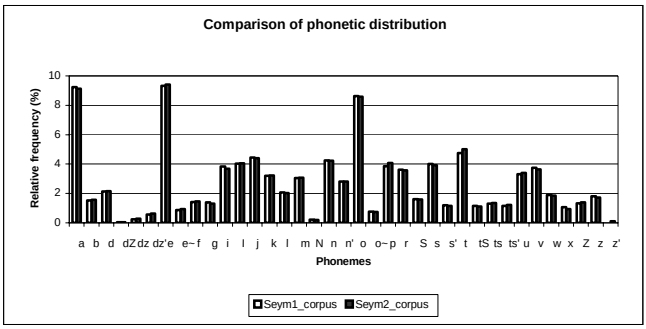
Similarly, the results show that the frequency of phonetic distribution is independent of the domain as far as the phonemes are concerned. It was decided that the preliminary selection will be based on parliamentary statements

5. Balancing the corpora

For balancing the parliamentary statements sub-corpora (about 22000 sentences each) were used . CorpusCrt program was used as a corpus balancing tool for sentence selection. The following criteria were used:

- The minimum phonetic length of a sentence is 30 phonemes;
- The maximum phonetic length of a sentence is 80 phonemes;
- The output corpus should contain 2500 sentences;
- Each phoneme should occur at least 40 times in the corpus;
- Each diphone should occur at least 4 times in the corpus;
- Each triphone should occur at least 3 times (this requirement is only possible for most frequent triphones)

These requirements were inputted to the greedy algorithm program (CorpusCrt) and twelve different versions of balanced corpora with 2500 sentences each have been created. Picture 4 shows the phonetic distribution from two randomly selected output corpora.



Picture 4. *Comparison of phonetic distribution between two random output corpora.*

A corpus of that size corresponds roughly to six hours of recordings and is considered to be sufficient for a corpus based speech synthesis system. A database generated this way provides about 1100 full diphone coverage. In the case of triphones the resulting text covers most frequent Polish triphones.

6. Two step corpus balancing

The first balancing phase described above resulted in 12 corpora, each of them balanced according to the criteria in 5. The next step involved joining these balanced corpora and balancing them once more.

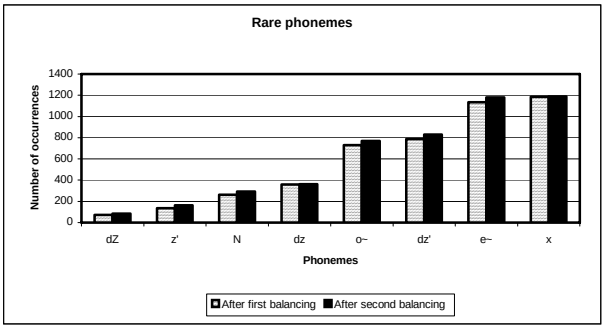
This step was motivated by attempt to optimise the frequency of each of the units (phonemes, diphones, triphones). Additionally, the number of rare phonemes is expected to rise proportionally to the size of the corpus.

Conforming our hypothesis, all unit coverage as well as occurrences of rare units has increased. For example, phoneme /dZ/ occurring on average 55 times in each sub-corpus, is after second balancing present 87 times (c.f. Picture 5). The advantages of increasing the number of rare phonemes has been studied by Beutnagel & Conkie [1]. They report that rare units are often preferred in their selection synthesis system and by including rare units in their database the quality of synthesis highly increased.

Here is the summary of acoustic optimisation changes after second balancing. The second iteration of CorpusCrt resulted in:

- longer sentence (58.3916 vs. 59.3256 phonemes);
- bigger overall phoneme coverage (145979 vs. 148314);
- greater average phoneme frequency (3945.38 vs. 4008.49).
- In the case of diphones 2nd balancing resulted in:
- increased diphone number (148479 vs. 150814)
- reduced number of diphones appearing less than four times from 175 to 68;

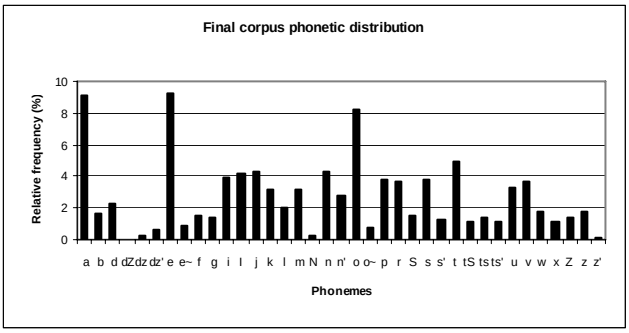
- increased number of different diphones from 1096 to 1196.
- The same process for triphones resulted in:
- increased triphone number (from 145979 to 148314);
- increased number of different diphones from 11524 to 13832.



Picture 5. *Comparison of occurrences of rare phonemes in 1st and 2nd balancing phase.*

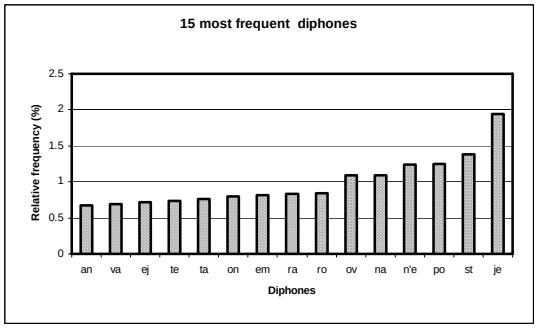
7. Results

The tree step sentence selection process has resulted in a final corpus of 2500 sentences. All are taken from an initial corpora of parliamentary statements. Picture 6 shows its phonetic distribution.



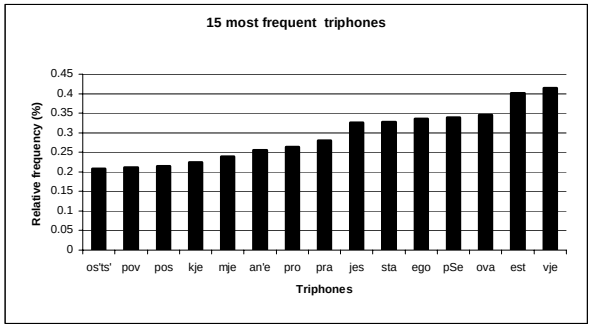
Picture 6. *Phonetic distribution in the final corpus.*

Picture 7 shows 15 most frequent diphones in our final corpus. They represent 14,8% of all diphones present in the corpus.



Picture 7. *15 most frequent diphones in the final corpus.*

Picture 8. shows 15 most frequent triphones, representing 4,4 % of all triphones in the selected corpus



Picture 8. *15 most frequent triphones in the final corpus.*

The final corpus contains 15776 different triphones.

7. Future work

The next step.will be the manual graphemic and phonetic correction of the corpus. It will be recorded in a few months and then automatic segmentation will be prepared.

8. Conclusion

In this study, we have presented the process of creating and optimizing a corpus for Polish unit selection speech synthesis. We have shown how a two step corpus balancing process results in better coverage of rare phonemes, diphones and triphones.

Sentences selected with this method will have to be manually verified in order to eliminate any markers, abbreviations acronyms which were not expanded in initial pre-processing. Manual correction of phonetic and graphemic transcription has been made. This will be followed by recording the selected sentences by a Polish voice talent. The ultimate goal of the project is the creation of unit selection speech synthesis system for Polish.

References

[1]Beutnagel, M. and Conkie, A. “Interaction of Units in a Unit Selection database”, Proc. Of Eurospeech, Budapest Hungary, 1999, p.1063-1066.

[2]Black, A. W. and Lenzo, K. "Optimal Utterance Selection for Unit Selection Speech Synthesis Databases", International Journal of Speech Technology, Vol. 6, p. 357--363, 2003.

[3]Black, A. W. and Lenzo, K. "Optimal Utterance Selection for Unit Selection Synthesis", In 4th ISCA Tutorial and Research Workshop on Speech Synthesis, Atholl Palace Hotel, Scotland, 2001.

[4]Bozkurt, B., Ozturk, O., Dutoit, T. "Text Design for TTS Speech Corpus Building using a Modified Greedy Selection". Proc. 5th Euro. Conf. on Speech Communication and Technology (EUROSPEECH-97), pp. 227--280, Rhodes, Greece, 1997.

[5]Kishore S. P., Black, A. W," Unit Size in Unit Selection Speech Synthesis", Eurospeech, Geneva 2003.

[6]Klabbers Esther, Stoeber Karlheinz, "Creation of Speech Corpora for the Multilingual Bonn Open Synthesis System"

[7]Oliver, D. "Polish Text to Speech synthesis system", MSc Thesis, Edinburgh University, Edinburgh, 1998

[8]Szklanny K. "Preparing the Polish diphone database for speech synthesis in MBROLA. "50. Otwarte Seminarium z Akustyki Szczyrk, Poland, 2003

[9]Van Santen, Jan P. H., and Adam L. Buchsbaum. "Methods for Optimal Text Selection", Proc. 5th Euro. Conf. on Speech Communication and Technology (EUROSPEECH-97), v. 2, pp. 553--6, Rhodes, Greece, 1997.

[10]Villasenor-Pineda, L., Montes-y-Gómez M., Pérez-Coutino, M. A.,Vaufreydaz D. "A Corpus Balancing Method for Language Model Construction", Computational Linguistics and Intelligent Text Processing, 4th International Conference, CICLing 2003, pp. 393--401, Mexico City, Mexico, 2003.