Proceedings of the Federated Conference on Computer Science and Information Systems pp. 1263–1267
DOI: 10.15439/2017F395
ISSN 2300-5963 ACSIS, Vol. 11

Implementation and verification of speech database for unit selection speech synthesis

Krzysztof Szklanny, Sebastian Koszuta
Polish-Japanese Academy of Information Technology, Multimedia Department, Koszykowa 86, Warsaw, Poland
Email: {kszklanny, s7127}@pjwstk.edu.pl

***Abstract*— The main aim of this study was to prepare a new speech database for the purpose of unit selection speech synthesis. The object was to design a database with improved parameters compared with the existing database [1], making use of the theses proved in studies [2]-[4]. The quality of the corpus, a selection of the suitable speaker, and the quality of the speech database are all crucially important for the quality of synthesized speech. The considerably larger text corpora used in the study as well as the broader multiple balancing of the database yielded a greater number of varied acoustic units. For the purpose of the recording, one voice talent was selected from among a group of 30 professional speakers. The next stage involved database segmentation. The resultant database was then verified with a prototype speech synthesizer. The quality of the synthetic speech was compared to that of synthetic speech obtained in other Polish unit selection speech synthesis systems. Consequently, the end result proved to be better than the one obtained in the previous study [4]. The database had been supplemented and extended, significantly enhancing the quality of synthesized speech.**

I. Introduction

UNIT selection speech synthesis remains an effective and popular method of concatenative synthesis, yielding speech which is closest to natural sounding human speech. The quality of synthesized speech depends on a number of factors. First and foremost, it is essential to create a comprehensive speech database which will form the core of the system. The database should comprise a variety of acoustic units (phonemes, diphones, syllables) produced in a range of different contexts, of different occurrence and length.

The first stage in the creation of speech database is the construction of a balanced corpus. This process involves a selection, from a large text database, of a number of sentences which best meet the input criteria. The larger the database, the more likely it is that the selected sentences will meet the set criteria. However, a larger corpus also means a greater computer processing capacity necessary to synthesize a single sentence. What is crucial is a proper balancing that will ensure an optimal database size while maintaining the right proportion of acoustic units characteristic of a particular language. The speech corpus is built in a semi-automatic way and then corrected manually. The manual part of the designing process is implemented in restricted domain speech synthesis such as the speaking clock and train departure announcements, and restricted speech recognition systems. The process is automated with the use of tools based on a greedy algorithm [5].

Another important aspect involves a careful selection of the speaker who will record the corpus. The speaker is usually voted on by experts, while an online questionnaire is often used to speed up the selection process. The recordings are made in a recording studio during a number of sessions, each several hours long. Each consecutive session is preceded by a hearing of the previously recorded material in order to establish a consistent volume, tone of voice, way of speaking, etc.

The final stage in the construction of speech database, following the recordings, is the appropriate labeling and segmentation. The segmentation of the database is carried out automatically with the use of statistical models, or heuristic methods, such as neural networks. Such a database should then be verified for the accuracy of the alignment of the defined boundaries of acoustic units.

The aim of this study was to design a new speech database with improved parameters. To this end, theses proved in [2]-[4] were used. The quality of the corpus, the selection of the right speaker and the quality of the database have a considerable influence on the quality of synthesized speech. The completed database was verified in a prototype synthesis engine.

II. Methods

A. Designing the speech database

The database was created with three corpora: no. 1 - a normalized collection of parliamentary speeches, stenographic records from select committee sessions, and extracts from IT e-books of 600MB (equivalent to 5 million sentences); no. 2 - subtitles for three feature films, i.e. Q. Tarantino's 1994 'Pulp Fiction', S. Kubrick's 1987 'Full Metal Jacket' and K. Smith's 1994 'Clerks', containing 4300 utterances; no. 3 - a corpus of 2150 sentences which served as a basis for the creation of the corpus-based speech synthesis [1],[4]. This corpus was based on a 300 MB text file containing, among others, a selection of parliamentary

This work was partially supported by the Research Centre of the Polish-Japanese Academy of Information Technology, supported by the Ministry of Science and Higher Education in Poland

speeches. It underwent multiple balancing (complying with the criteria outlined in section 2.3) and was supplemented with low frequency phonemes. The final corpus includes 1196 different diphones and 11524 triphones [1],[4].

Corpus no. 1 was subdivided into 250 files, each containing 20,000 sentences, of which 16 sub-corpora were randomly selected for further processing. Such a division makes data processing more efficient. In the final stage of the balancing, corpora no. 2 and no. 3 were used to expand the newly designed corpus. Findings presented in [2] indicate that multiple balancing helps to make the corpus more representative, thereby enhancing the quality of speech synthesis.

B. Phonetic transcription

Phonetic transcription makes it possible to convert orthographic text into phonetic script. This is done by means of a special phonetic alphabet, such as PL-SAMPA [6].

The automatic phonetic transcription was generated with the help of software available as part of the Clarin project [7]. The application operates within a rule-based system. The diphone and triphone transcriptions were generated in Perl.

C. Multiple balancing

The CorpusCrt program is an implementation of Greedy algorithm [8]. It was used as a balancing tool for sentence selection. Each of the 16 sub-corpora was balanced according to the following criteria:

- Each sentence should contain a minimum of 16 phonemes;
- Each sentence should contain a maximum of 80 phonemes;
- Each phoneme should occur at least 40 times in the entire corpus;
- Each diphone should occur at least 4 times in the entire corpus;
- Each triphone should occur at least 3 times in the entire corpus (due to the large number of possible triphones, this particular criterion could only be met for 400 most frequently used triphones in the Polish language);
- The output corpus should contain 2500 sentences.

TABLE I. PERCENT FREQUENCY DISTRIBUTION OF LOWEST-FREQUENCY POLISH PHONEMES IN A RANDOMLY SELECTED SUB-CORPUS BEFORE AND AFTER THE INITIAL BALANCING

Phoneme	Before balancing	After balancing
dZ	0.01%	0.02%
z'	0.10%	0.16%
N	0.20%	0.17%
dz	0.31%	0.36%
o~	0.59%	0.77%
dz'	0.76%	0.78%
X	0.79%	0.87%
ts'	0.83%	0.94%
e~	0.78%	1.09%

Table I shows a percent frequency distribution of lowest frequency polish phonemes in a randomly selected sub-corpus before and after the initial balancing.

The aim of the second balancing was to create one corpus that would include the phonetically richest sentences from the 16 already existent sub-corpora. The sub-corpora were first merged into a file of 40,000 utterances which, when balanced, yielded a corpus of 2,500 sentences. The result was a richer coverage of acoustic units in comparison to each of the separate sub-corpora.

1) Merging with the corpus assigned for unit-selection speech synthesis

The resultant corpus was then merged with corpus no. 3 and balanced to 2,500 sentences. The number of low-frequency phonemes (DZ, z', N, o~, e~) increased from 148 879 to 149 635.

It was essential that the corpus contained a wide range of prosodic contexts for the different phonetic components. Therefore, it was subsequently supplemented with prosodic features from corpus no. 2. This involved using all the interrogative and exclamatory sentences. The corpus was then balanced to yield two corpora of 50 sentences each. The first one contained interrogative sentences, while the other contained exclamatory sentences. These corpora were then concatenated with the main corpus (without further balancing). Previous findings indicate [2] that it is possible to reduce the size of a corpus. In the final balancing, the corpus was reduced to 2,150 sentences, with the assumption that a corpus must contain a minimum of 15,000 triphones while the number of diphones must remain unchanged. The average length of a sentence in the corpus is that of 63.93, whereas the total number of phonemes is 128,169. The corpus contains 1279 different diphones and 15,087 different triphones. Table II shows data concerning the number of acoustic units depending on the size of a corpus. Fig. 1 shows a percent frequency distribution of phonemes in the final corpus.

TABLE II. NUMBER OF ACOUSTIC UNITS AFTER CORPUS SIZE REDUCTION WHICH SERVED AS A BASIS FOR THE SELECTION OF THE FINAL CORPUS

No. of sentences	**2600**	**2400**	**2200**	**2150**	**2100**
No. of diphones	1279	1279	1279	1279	1279
No. of triphones	15869	15615	15218	15078	14979
No. of triphones < 3	8379	8387	8285	8228	8189
No. of diphones < 5	165	184	199	199	203

D. Speaker selection and recordings

The speaker was selected on the basis of recorded voice samples collected from 30 candidates. Each candidate was a voice talent. The objective was to find a speaker with a strong steady voice. The voice assessment was carried out by eight voice analysis experts, who chose a female voice.

The recordings were conducted in the recording studio of the Polish-Japanese Institute of Information Technology, Warsaw (now Polish-Japanese Academy of Information

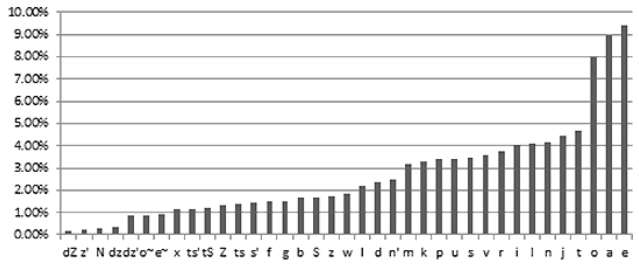

Fig.1: Percent frequency distribution of phonemes in the final corpus

Technology), using an Audio-Technica AT2020 microphone with a pop filter. The signal was recorded in the AIFF format with a 48 kHz sampling frequency and a 24 bit resolution, using the audio Focusrite Scarlett 2i4 interface.

The corpus was recorded during 15 two-hour sessions. Each prompt was recorded as a separate file. After each session, the files were exported in the WAV format with file names corresponding to the prompt numbers in the corpus. The recordings were then checked for distortions and external noises as well as mistakes made by the speaker. 480 prompts were re-recorded.

E. Segmentation

The automatic segmentation was carried out with a program based on the Kaldi project [9]. Kaldi is an open source speech recognition toolkit, written in C++. The segmentation was based on the 'forced alignment' technique, which involves matching phoneme boundaries on the basis of a file containing phonetic transcription. First, the program creates an FST graph whose states correspond to the consecutive segmental phonemes of the analyzed phrase. Following that, a sequence of states with set boundaries is assigned for recording, by means of the Viterbi algorithm. The phonetic transcription for the segmentation was performed on the basis of an orthographic transcription using a Polish language dictionary with SAMPA transcriptions. The transcription of foreign words and proper nouns was performed manually [10].

III. VERIFICATION OF THE SPEECH DATABASE

To examine the quality of the speech database and to verify the quality of the segmentation, a prototype speech synthesizer, written in Java, was used to conduct a series of tests. The program does not contain the NLP module but allows a preliminary evaluation of the quality of the corpus. It facilitates unit selection using three different algorithms: 'Random', 'Forward' and 'Viterbi' (the so-called Viterbi algorithm) [11]. These algorithms are responsible for the way acoustic units are selected from the database. The main criterion that is taken into account in the selection of acoustic units is their direct neighborhood in the database, which reduces the likelihood of the occurrence of artifacts, such as energy discontinuity, which render synthesized speech artificial. The similarity of F_0 at the boundaries of concatenated units is also taken into account.

The 'Random' algorithm randomly selects acoustic units that match the phonetic transcription, without cost function. Its application is the least effective of all the three algorithms.

'Forward' and 'Viterbi' are more advanced algorithms which make it possible to use cost function for the comparison of hypotheses. In unit selection speech synthesis, a hypothesis is a sequence of acoustic units selected from the database which, having been concatenated, produce a phrase that is to be synthesized. The object is to select a sequence that will produce the most natural sounding speech. These two algorithms are similar and yield similar results. The Viterbi algorithm was chosen for the testing process. The searching process is based on the trellis of all the candidates which is formed by the paths between them. The Viterbi algorithm searches the trellis from left to right, calculating partial costs, which is the sum total of the sequences of the cost function. The optimum path with the lowest cost is then chosen.

The prototype synthesizer utilizes MLF files (with diphone boundaries in the corpus), WAV sound files (with recorded prompts), and files containing data about F_0 for each of the prompts. The text to be synthesized is provided in the form of a phonetic transcription.

IV. RESULTS

A Mean Opinion Score (MOS) test was designed to check the quality of the synthesizer. MOS is a subjective measure for audio and video quality evaluation. In the test, subjects are administered audio or video samples, after which they give their subjective opinion using the following five-point scale: 1 – bad, 2 – poor, 3 – fair, 4 – good, 5 – excellent.
The MOS is expressed as the arithmetic mean of all the collected ratings. MOS is also recommended as a method for evaluating the quality of synthesized speech [12]. To assess the quality of the voice a special website with an online questionnaire was designed, which served as an anonymous tool for evaluating speech samples on the five-point scale. The test involved 14 individuals who were familiar with issues related to speech synthesis, phonetics of the Polish language and phonetic transcription, and who were also well-informed about natural language processing. The test was divided into three parts. The first five recorded sentences were used to judge the quality of lector voice; the samples were then used to generate another five resynthesized sentences; the third part of the test involved sentences synthesized in the prototype speech synthesizer. Long, phonetically rich sentences were selected to this end.The first part of the test received the average score of 4.3, which indicates that the speaker's voice was rated high by the experts. The speaker's voice rating reflects the respondents' opinion concerning the potential effectiveness of the future synthesizer. It is the maximum score that the best synthesizer could receive. Resynthesis of sentences

inevitably involves a decrease in their quality. In the test, the quality of the synthesis received an average opinion score of 3.41, which is a good result. The third part of the test received an opinion score of 2.07.

V. Discussion

It would be worthwhile to compare the obtained results with the commercial and non-commercial systems functioning in Poland, taking into account the evaluation of the quality of the entire system and not merely the speech database.

The first Polish system for unit selection speech synthesis was BOSS, which was created as part of a collaborative research project between Adam Mickiewicz University, Poznan and IKP (Institut für Kommunikationsforschung und Phonetik) in Bonn [13]-[15]. The speech database consists of approximately 115 minutes of audio material read by a professional speaker, recorded during several sessions and supervised by an expert phonetician. The database is subdivided into six parts. The first part consists of phrases with most frequent consonant structures, where 258 consonant clusters of various types are used. The second part consists of all Polish diphones realised in 92 grammatically correct but semantically nonsense phrases. The third part consists of 664 phrases with CVC triphones (consonant-vowel-consonant, in non-sonorant voiced context and with various intonation patterns). The fourth part consists of 985 phrases, each made up of 6 to 14 syllables. The fifth part consists of 1109 sentences made up of 6000 most frequent vocabulary items. The sixth part consists of 15-minute long prose passages and newspaper articles [16]. The database was implemented in the Bonn Speech Synthesis System. A three-part MOS test was conducted for the designed system: the first part involved common utterances – 25 sentences and phrases created especially for the purpose, mostly using the top high frequency vocabulary items from a large vocabulary newspaper frequency list, and conversational utterances; the second part comprised 25 typical Polish conversational phrases, dialogue phrases, short expressions and natural utterances; the third part comprised a reference set, i.e. 24 original recordings of the speaker reading short utterances. The speaker's voice received an opinion score of 4.6, whereas the speech synthesis system received a score of 3.39. Further experiments, which involved manual correction of the speech database while focusing on duration weighting, increased the MOS opinion score to 3.62 [17] for the speech synthesis system. The quality of synthesized speech based on automatically segmented database received an overall score of 2.44. This result covers re-synthesized sentences from the corpus, sentences with high frequency vocabulary items and words that are 'difficult' for the synthesizer, i.e. phonetically rich items.

However, the quality rating for difficult sentences, i.e. sentences similar to those used for testing the original database, was 1.70, which then rose to 1.71 following a manual correction of the segmentation. Unfortunately, the publication [17],[18] does not present the tested sentences, which could be used to evaluate the quality of the database.

IVONA, a commercial system for unit selection speech synthesis, was created by IVOSOFTWARE (now Amazon). In the Blizzard Challenge 2006, the system received the following opinion scores: 4.66 for the speaker, and 3.69 for the quality of synthesis with an ATR database [19],[20]. In 2007, the scores were 4.70 and 3.90 respectively, using the same database. In the 2009 Blizzard Challenge, IVONA received 4.90 for the speaker and 4.00 for the quality of the synthesis, with an EH1 database [21]. The presented data concerns speech synthesis for the English language. However, no publication presenting MOS results for the Polish language is available.

Tests were also conducted for the original synthetic speech system that was developed in the Festival meta system [22]. These were carried out following work on a speech synthesizer [4]. 28 experts were involved in the tests, and the average MOS result for the speaker's voice was 4.60. The experts assessed the quality of the resynthesis at 3.79, which is a good result. Sentence synthesis with the best cost function, optimized with an evolutionary algorithm, received an opinion score of 2.71, the worst cost function 1.97, and the default cost function 2.19. These results are worse than those obtained for the other speech synthesis systems. However, it must be noted that the basic problem stemmed from the construction of a database recorded by a non-professional speaker. The utterances exhibited considerable F_0 fluctuations, which in turn affected the right selection of appropriate acoustic units. Despite this, the synthesis in the complete speech synthesizer with a default cost function received a score similar to that of a new database that was tested in the prototype synthesizer (2.11 vs. 2.07), even though the segmentation quality did not undergo manual correction. Compared with the BOSS system, this result is better for phonetically rich sentences.

When comparing the opinion scores of recorded samples and resynthesized samples, one can notice a significant discrepancy (0.88). This may indicate errors in the functioning of the prototype synthesizer and/or an incorrect phonetic transcription used in the selection of acoustic units for speech synthesis. Other reasons may include the presence of elements of acoustic units which appear in synthesized sentences as a result of automatic segmentation. This problem can be eliminated by manual correction. One of the methods is described in [23]. This kind of correction, as well as improvements made to the prototype speech synthesizer, will ensure a higher opinion score. Criteria applied in previous studies [23] will still be used in order to detect durational outliers. These include phonemes of abnormal duration, zero crossing errors, plosive phonemes and other distortions.

The construction of the new speech database made it possible to eliminate the errors which the author encountered when designing the previous database. These involved the quality of the speaker's voice, including the

excessively fast speech delivery, and considerable F_0 fluctuations in sentences. What was also eliminated was the errors that occurred at the corpus building stage. The corpus was extended to include utterances from everyday speech, which should improve the quality of synthesized sentences in this area.

VI. Conclusions

When designing the speech database, the author drew on the experience gained during the implementation of the unit selection speech synthesis. The corpus was supplemented and extended, and the recordings were made by a professional speaker selected by means of tests, which is crucial for the quality of synthetically generated speech. The database created for previous studies was recorded by a semi-professional speaker.

Despite the fact that manual segmentation correction was not performed, the results obtained in a MOS test were similar to those of a manually corrected database (2.07 vs. 2.18), and its opinion score for phonetically rich sentences was higher than that for the BOSS database (2.07 vs. 1.70).

What it means is that the elimination of other errors during the implementation of the new speech synthesis system will make it possible to achieve a higher quality of synthesized speech, comparable to that of the BOSS and IVONA synthetic speech systems. The next stage of the research will be to incorporate the database into the existent multimodal speech synthesis. We also plan to verify and place the database in compliance with the ECESS standards and to arrange for the database to be validated by an independent institution, such as ELDA [24].

Acknowledgment

The author would like to thank Danijel Koržinek for his help with the implementation of the prototype speech synthesizer and Prof. Krzysztof Marasek for his help in finding professional speaker.

References

[1] D. Oliver, K. Szklanny, (2006). Creation and analysis of a Polish speech database for use in unit selection synthesis. In *LREC-2006: Fifth International Conference on Language Resources and Evaluation*.

[2] K. Szklanny „Optymalizacja funkcji kosztu w korpusowej syntezie mowy polskiej". Diss. Polsko-Japońska Wyższa Szkoła Technik Komputerowych, 2009.

[3] K. Szklanny "System Korpusowej Syntezy Mowy Dla Języka Polskiego." XI International PhD Workshop OWD 2009, 17–20 October 2009

[4] K. Szklanny (2014). "Multimodal Speech Synthesis for Polish Language. In Man-Machine Interactions 3 (pp. 325-333). Springer International Publishing." DOI: 10.1007/978-3-319-02309-0_35

[5] B. Bozkurt, T. Dutoit, O. Ozturk: Text Design For TTS Speech Corpus Building Using A Modified Greedy Selection, Proc. Eurospeech, Geneva 2003, pp 277-280.

[6] J.C. Wells (1997) SAMPA computer readable phonetic alphabet, in Gibbon, D., Moore, R. and Winski, R. (eds.), 1997. Handbook of Standards and Resources for Spoken Language Systems. Berlin and New York: Mouton de Gruyter. Part IV, section B.

[7] D. Koržinek, K. Marasek, Ł. Brocki, 2016, Polish Speech Services, CLARIN-PL digital repository, http://hdl.handle.net/11321/296.

[8] A. S. Bailador. 1998. CorpusCrt. Technical report, Polytechnic University of Catalonia (UPC).

[9] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, K. Vesely, *The Kaldi Speech Recognition Toolkit*

[10] Marasek, K., Koržinek, D. and Brocki, Ł. (2015). System for Automatic Transcription of Sessions of the Polish Senate. *Archives of Acoustics*, 39(4). DOI: https://doi.org/10.2478/aoa-2014-0054

[11] A. J. Viterbi (1967) Error bounds for convolutional codes and an asymptotically optimal decoding algorithm, IEEE Transactions on Information Processing, 13:260-269

[12] ITU-T reccomendation no P.85 (https://www.itu.int/rec/T-REC-P.85-199406-I/en).

[13] E. Klabbers, K. Stöber, R. Veldhuis, P. Wagner, S. Breuer (2001 B) Speech synthesis development made easy: The Bonn Open Synthesis System, Eurospeech 2001, Aalborg,

[14] G. Demenko, K. Klessa, M. Szymański, J. Bachan (2007) The design of Polish speech corpora for speech synthesis in BOSS system, Mat.XII Sympozjum Podstawowe Problemy Energoelektroniki, Elektromechaniki i Mechatroniki (PPEEm'2007), Wisła, Poland, pp. 253-258.

[15] G. Demenko, A. Wagner (2007) Prosody annotation for unit selection text-to-speech synthesis, Archives of acoustics, 32(1):.25-40

[16] G. Demenko, J. Bachan, B. Möbius, K. Klessa, M. Szymański, S. Grocholewski, (2008). Development and evaluation of Polish speech corpus for unit selection speech synthesis systems. In *Ninth Annual Conference of the International Speech Communication Association*.

[17] M. Szymański, K. Kleesa, and G. Demenko. "Optimization of unit selection speech synthesis." Proceedings of 17th International Congress of Phonetic Sciences (ICPhS 2011). 2011.

[18] G. Demenko, K. Klessa, M. Szymański, S. Breuer, & W. Hess, (2010). Polish unit selection speech synthesis with BOSS: extensions and speech corpora. *International Journal of Speech Technology*, *13*(2), 85-99. DOI: 10.1007/s10772-010-9071-3

[19] M. Kaszczuk, L. Osowski. "Evaluating Ivona speech synthesis system for Blizzard Challenge 2006." *Blizzard Workshop, Pittsburgh*. 2006.

[20] M. Kaszczuk, L. Osowski. "The IVO Software Blizzard 2007 Entry: Improving Ivona Speech Synthesis System." Sixth ISCA Workshop on Speech Synthesis, Bonn. 2007.

[21] M. Kaszczuk, L Osowski. "The IVO software Blizzard Challenge 2009 entry: Improving IVONA text-to-speech." *Blizzard Challenge Workshop*. 2009.

[22] R. Clark, K. Richmond, & S. King, (2007). Multisyn: Open-domain unit selection for the Festival speech synthesis system. Speech Communication, 49(4), 317-330. http://dx.doi.org/10.1016/j.specom.2007.01.014

[23] K. Szklanny, M. Wojtowski, (2008, May). Automatic segmentation quality improvement for realization of unit selection speech synthesis. In 2008 Conference on Human System Interactions (pp. 251-256). IEEE, DOI: 10.1109/HSI.2008.4581443.

[24] ELDA: Evaluations and Language resources Distribution Agency. Online: http://www.elda.org/, accessed on 21 April 2017.