Design of an Efficient Corpus for High-Quality Unit Selection TTS for Bulgarian

Aimilios Chalamandaris, Pirros Tsiakoulis, Spyros Raptis, Sotiris Karabetsos

Institute for Language and Speech Processing – Athena Research Centre
Artemidos 6 & Epidavrou, 15125, Athens, Greece
{achalam, ptsiak, spy, sotoskar}@ilsp.gr

Abstract

In this paper we present the process of designing an efficient speech corpus for the first unit selection speech synthesis system for Bulgarian[1] along with some preliminary results. As the initial corpus is a crucial factor for the quality delivered by the Text-to-Speech system, special effort has been given to design a complete and efficient corpus for use in a unit selection TTS system. The targeted domain of the TTS system and hence of the corpus is the news reports, and although it is a restricted one, it is characterized by an unlimited vocabulary. The paper focuses on issues regarding the design of an optimal corpus for such a framework and the ideas on which our approach was based on. A novel multi-stage approach is presented for efficient corpus design, with special attention given to language and speaker dependent issues, as they affect the entire process. The paper concludes with the presentation of our results and the evaluation experiments, which provide clear evidence of the quality level achieved.

1. Introduction

Text-to-speech synthesis systems convert textual input into synthetic voice signals. By offering the promise of natural and intuitive human-computer interaction, text to speech synthesis systems have gained considerable attention in the course of their development, both at the side of their design and implementation and at the side of their applications. In recent years, text-to-speech (TTS) systems have shown a significant improvement as far as the quality of the synthetic speech is concerned. This evolution has been sparked mainly by the fact that they rely more and more on data-driven, statistical modeling of the speech, with time or frequency domain algorithms for signal manipulation (Moulines 1990, Schroeter 2008, Möbius 2000, Nagy 2005). Current approaches aim to model the speech of a given human speaker, based on recordings of his voice, and deliver synthetic speech which manages to capture its spectral and prosodic characteristics, such as voice timbre, pitch and durations. The key idea of unit selection speech synthesis (Hunt, 1996), which currently constitutes the cornerstone of most of the state-of-the-art TTS systems worldwide, is to use an entire speech corpus as the acoustic inventory and to select at run-time from this corpus different acoustic units that match better according to a metric, so as to capture the characteristics of a targeted synthetic speech. Although recent parametric approaches for speech synthesis have emerged, such as HMM-based speech synthesis (Tokuda 2000) aiming to model the features of one's voice through Hidden Markov Models with significantly good and promising results, the Unit Selection method provides a wide framework for speech synthesis, which has been advocated to best meet the needs of limited domains.

It is a common knowledge that the quality of the synthetic speech provided by a corpus based TTS system highly depends on the quality of its acoustic unit inventory. The most important factor affecting the quality of this inventory is the initial text corpus, namely the set of sentences that the speaker will have to utter and record, in order to create the initial data for the acoustic inventory.

There have been several proposals on how to create a corpus for a TTS system (Black 2003, Bozkurt 2003, Iida 2001, Matousek 2001). The cornerstone of almost all proposals is the use of greedy algorithms in order to select a sub-corpus from an initial large corpus pool, which would fulfill satisfactorily several different requirements. These requirements can be as simple as word coverage (Iida 2003, Lewis 1999), or more complex such as phonetic or prosodic coverage, or the coverage of other language-dependent or domain-dependent parameters (Lambert 2006). Our methodology attempts to deal with different types of parameters in order to design an efficient spoken corpus for a relatively general-domain TTS system for Bulgarian. Although the targeted domain, namely the news reports domain is restricted, it can also be regarded as relatively generic because of the wide range of topics it covers.

The main objectives of our approach were two: (a) to achieve sufficient coverage of the significant language-dependant phenomena identified and (b) to ensure consistently good performance during synthesis. More specifically, the first objective can be projected to the following set of complementary goals: (i) phonetic coverage, (ii) prosodic coverage, and (iii) controlled redundancy, while the second objective aims to compensate for problems that are relevant either to the speaker's voice characteristics or the actual recordings and cause the TTS system to present inconsistent performance.

2. Corpus Design Strategies

In a unit selection TTS system, the corpus design problem can be regarded as a set coverage problem. The target set *C* is the set of units to be covered. Each sentence in the corpus is also a set of units, and the corpus selection problem consists in finding a minimum size set of sentences which will contain all the units defined in the set *C*.

[1] The unit selection TTS system for Bulgarian mentioned in this paper is developed at the Institute for Language and Speech Processing ("Athena" Research Centre), in Greece (Raptis, 2009).

According to the domain specifics and needs of the application field, the nature and number of units can vary from simply few words to many thousands of phonemes, diphones or longer units. In the extreme case of very limited domain (Iida 2001, Yi 1998), it has been shown that if a database is deliberately tailored to the intended application, the TTS system can provide robustly high quality synthetic speech. In such cases, the process of corpus design is to include in the sentences at least one occurrence of each word in the domain, in each desired prosodic context (Matousek 2003). Manual selection or compilation of the sentences can often be adequate for such limited domains.

Different units allow for different strategies for corpus design. In the case of very limited domains, such as weather reporting, the units can be words or even phrases that can be reproduced during text to speech synthesis. Such approaches have been very popular for such domains (Kishore 2003), and it has been shown that they provide high quality synthetic speech. However, in the case of a more generic domain, or of a domain with unlimited vocabulary, even though it is practically restricted, such as ours, the units cannot be as long as words, since it is impossible to calculate efficient coverage (Schweitzer 2003).

Many researches use diphones as basic units for the unit selection and therefore they also use diphones in order to handle the corpus design problem. Others identify severe drawbacks when employing diphones for corpus design based on the theory of Large Number of Rare Events (LNRE) (Möbius 2003) and they suggest either modifying the searching algorithm using diphones, or using triphones as basic units (Bozkurt 2003).

In our approach the basic selection unit is the diphone, for each of which we employ a feature vector that sufficiently describes its contextual properties, as far as the TTS system is concerned. More details about the employed features are provided in section 3.

Utterance Selection Methods

The corpus design problem as defined in previous paragraphs can be regarded as a process for deriving a minimum size set of sentences which offer coverage for a target units set *C*. The target unit set *C* incorporates all target units necessary for the TTS system to deliver high quality synthetic speech. The common practice for designing such a corpus automatically is by employing a greedy algorithm (Franois 2002). The latter is an iterative technique for compiling a subset of sentences from a large set of sentences (corpus pool) in order to cover the largest unit space *C* with the smallest number of sentences. Prior to the selection, the corpus pool as well as the target set *C* must be well defined. Normally, the initial corpus pool is a set of sentences that well define the text style of the targeted domain of the TTS application. That is, if for example one would aim to create a corpus for sports reports, then the initial corpus pool should contain mainly texts from sports news and reports. As far as the target set *C* is concerned, it often consists of units that best describe the phonetic content of the targeted domain (Bozkurt 2003, Black 2003). Rational extension to this idea is the inclusion of other important parameters such as prosodic, stress, contextual information of the phonemes to be covered (Matousek 2003, Lambert 2005).

The greedy selection algorithm involves assigning costs to every sentence of the corpus pool according to the number of units that are in common with the target set *C*, and the number of units that are not in their intersection. At each iteration, the algorithm selects the sentence with the highest ranking according to the previous criterion, it removes it from the corpus pool, and it updates the target set *C* by removing the units that have been covered by this sentence. This process continues until a termination criterion is reached, such as maximum number of sentences or efficient coverage of the target set *C*. The main drawback of this technique is that if the number of factors defining a unit is large, that is, if the target set *C* is significantly large, then the produced corpus may be prohibitively large. Modified greedy algorithms which mainly aim to indirectly cluster the factors defining the units of the target set *C* have also been suggested and they have shown to work efficiently (Black 2003).

The main aspects, on which most strategies employing the greedy algorithm differentiate from each other, are the statistical properties of the coverage set. Often it is suggested that a phonetic distribution similar to the one of the corpus pool should be aimed to be achieved in order to better capture the acoustic properties of the initial large corpus, while in other cases, priority is given on the rarest unit classes in order to compensate for LNRE phenomena (Ozkurt 2003, Andersen 2003).

3. The Proposed Corpus Selection Method

It has been clear from our experiments and from other researches that the process of corpus design for a unit selection TTS is not trivial and it should be given special attention and effort. It has also been clear that the unit selection TTS systems suffer from limitations as far as the signal modification is concerned, and therefore their final quality is inherently dependent on the abundance and completeness of their database (Balestri 1999). Nevertheless, it is also obvious that the development of huge databases, aside from being a time and effort consuming process, does not necessarily guarantee proportionally to the size good results.

Our corpus design method aims to deliver an efficient corpus for a restricted but unlimited domain, with a special providence to identify and alleviate problems that would jeopardize consistency in the final quality. The main idea behind our algorithm is to define a process that will take advantage of available information about the targeted TTS system and, as a post-hoc process, it will be able to act complementarily to the unit selection algorithm's properties. By doing so, we cater for both the specifics of the domain we are targeting, as well as for the optimizing in such a way that it will behave optimally with our specific unit selection algorithm. Our results justify our approach and the underlying hypothesis.

The algorithm works in three stages:

1. From a domain specific extensively large corpus pool, we identify a set of sentences *S* which offers the maximum possible coverage of our unit target set *C* (as described below).
2. After the recording of the *S* set and the incorporation of it in the unit selection TTS system, we simulate the synthesis of the original

large corpus pool, in order to identify possible concatenation problems that are either speaker or sentence specific. By doing so, we automatically select an additional subset of sentences *S*' that present large number of problems in combination to our unit selection module.

3. In the final stage of our method we identify possible diphones that are missing from the initial large corpus pool and we manually insert them in short non-sense sentences compiling a new subset *S''*.

The final optimal corpus is then the union of the independently derived sets: $S_{Final} = S + S' + S''$
The stages described above deserve more explanation.

Stage 1: Selection of the *S* corpus set
Initially the large corpus set, serving as the corpus pool from which we automatically select the sentences, has to be collected and processed accordingly in order to become appropriate for this task. In order to break down a large text into a set of sentences two tasks have to be carried out initially: (i) *text normalization* and (ii) *sentence tokenization* (Schroeter 2008). Text normalization is responsible for the expansion of numerals, abbreviations and acronyms, as well as dates, addresses etc. It is a rather complex task, based mostly on heuristics and hybrid algorithms, in order to achieve disambiguation when necessary. The sentence tokenization process, although seems to be simple enough, it highly depends upon other text pre-processing modules and its level of complexity is often language-dependent. In order to phonetically transcribe every sentence, we used the grapheme to phoneme module we developed for Bulgarian (Raptis 2009).
From the phonetically transcribed corpus pool, where the prosodic characteristics of every unit are also provided by our prosody engine, we greedily select a subset of sentences that offers satisfactory coverage for our target unit set *C*. In order to define *C*, we constructed a contextual feature vector for diphone units that included key prosodic factors, such as word accent status, position in the utterance, distance from prosodic modifiers in the utterance etc. The prosodic factors employed here derive from our prosody engine incorporated into our TTS system, which is based on a data-driven prosody modeling approach. By using a greedy selection algorithm as described previously, we produce a set of sentences *S* which satisfactorily covers our defined target unit set.

Stage 2: Enhancement and fine tuning of the *S* corpus set
This stage of our approach provides a novel method for enhancing the speech corpus. During this stage, after the recording and processing of the sentences and their packaging into a database, ready to be used by our TTS system, we aim to identify problems during actual synthesis, which originate either from speaker-dependent factors, or from other factors, such as misaligned or even bad recordings. This process consists in synthesizing every sentence of the large corpus pool and identifying local and global maxima of the total cost function during the unit selection. By synthesis, we identify additional units that should also be covered by the final corpus, and through an iterative selection algorithm we select a set of sentences (*S'*) that optimally cover these units. Special care has been taken in order to maintain the recording conditions consistent throughout the entire process, especially because different recording sessions were carried out with long intervals between them.

Stage 3: Further enhancement with missing units
The final stage of our selection strategy although may seem trivial, is crucial for the better performance of the TTS system in different domains, or in difficult vocabularies, such as foreign words. This step of the selection process, leading to *S''*, practically aims to efficiently handle LNRE aspects as well, since whatever units are missing from the large corpus pool can be characterized as rare events. These units, even though they are very rare, they can affect the overall quality of the TTS system by fusing inconsistencies and mismatches in the synthesized speech.

4. Results

It was decided that the targeted domain would be the news reports, mainly for two reasons: (i) it is of the scope of the authors to develop a synthetic voice for news, and (ii) because it is both a restricted and unlimited domain and the usual neutral informative speaking style can also serve in other domains such as dialogues or speaking applications.
In order to shape the large corpus pool from which we would extract the optimal corpus, we collected the online news articles from different Bulgarian newspapers for a period of 12 months. The initial large corpus pool consisted of about 54 million words. After having performed text normalization and sentence tokenization onto the corpus set, we ended up with an initial corpus pool of about 4.15 million sentences. Sentences with foreign words, or extremely long or short ones, were discarded. Although other approaches explicitly decide to select from within sentences of reasonably short length, such as of maximum 10-12 words (Black 2003), we believe that longer sentences, although they might be difficult to pronounce or process, they incorporate "ingredients" that short sentences lack, such as more variable prosodic structure.
For creating the target unit set *C*, we employed the letter to sound module for the Bulgarian language we have developed in the framework of our TTS system and the respective prosody engine on the entire corpus (Raptis 2009), and we formed a feature vector for each unit containing the following parameters: (i) diphone type, (ii) acoustic context, (iii) prosodic cluster type and (iv) intra-prosodic relative position. We identified 204,514 unique units in the large corpus pool, which altogether consist our target unit set *C*. A custom greedy selection algorithm with a termination criterion of 4,000 sentences, selected a subset of sentences (*S*) from the large corpus pool, which offered not complete but sufficient coverage of the target unit set *C*. The selection algorithm was designed to pursue full coverage for all unique diphones and efficient unit coverage.
One of the major problems in such cases where a set of sentences is selected automatically with a dual criterion, to maximize the coverage and minimize the size, is the fact that often sentences with misspellings or other errors

are selected, since they provide coverage of rare acoustic events. In order to alleviate this problem, an additional mechanism has also been implemented that allowed us to review the rare instances in the resulted sentences, and observe the units for which each of them had been selected. During the manual correction of the sentences, if a correction would lead to the exclusion of a covered unit, then we simply removed the sentence from the set *S* and the system would suggest one or more other sentences that would compensate for all the units that the removed sentence was selected for. This process is necessary in order to remove any errors that could affect the final results at several intermediate stages. After this phase of the process, the corpus set *S* consisted of 4,083 sentences with 16.13 words on average.

The following two stages of the design process are carried out after the recordings and their processing for their incorporation into the TTS system. Hence, after the completion of the recordings and the corresponding database, we synthesized every sentence from the large corpus pool and we identified automatically the most problematic units, as far as the synthesis process is concerned. This was carried out automatically by identifying the local maxima in the unit selection total cost function for every synthesized sentence. With this process after a necessary clustering of the problematic units, we collected a set of units that were necessary to be additionally covered by the recordings. Again, through the means of the greedy selection algorithm and with a termination criterion in the number of the selected sentences, we automatically selected 1000 additional sentences, which would be included in a following recording session, by the same speaker. It is worth noting here, that special attention was given to carefully profile the recording settings for every session in order to ensure no deviations in the recorded speech.

The final stage of the selection process included the research and manual enrichment of the optimal sentence set with sentences containing possible missing diphones that can be met in the Bulgarian spoken language, even if they are only necessary for the pronunciation of foreign words with the Bulgarian phonetic alphabet. The additional sentences were non-sense sentences, of short length, and were produced manually by the concatenation of words which ensured the utterance of such diphones. This stage resulted in an additional set of 32 sentences. In the following table one can see the properties of the selected corpus.

Corpus Design Stage	# Sentences	Diphone Coverage	Unit Coverage
1st Stage	4,083	96%	61.6%
2nd Stage	5,083	96%	65.6%
3rd Stage	5,115	100%	66.1%

Table 1: The properties of the resulted corpus during the corpus design process.

5. Experimental Evaluation

To assess the effect of the Bulgarian speech synthesis system, a set of acoustic experiments was performed. The experiments targeted different dimensions of the quality, covering naturalness, intelligibility and speech flow. A final set of questions was used to capture the participants' opinion regarding the appropriateness of the synthesis system for different application areas. Finally, the listeners were given the option to provide free-text feedback.

The subjects were 30 native Bulgarian speakers participated, 10 of which had a background in linguistics or previous experience related to the subject and, for the purposes of these experiments, where considered as a distinct group.

The results of the experiments are illustrated in the following tables. A more detailed description of the experiments and analysis of the respective results can be found in (Raptis 2009).

		Experiment 1 (sentence-level)		
		Naturalness	Ease of listening	Articulation
Non-expert listeners	**MOS**	3,53	4,41	4,13
	STD	0,96	0,66	0,77
"Expert" listeners	**MOS**	3,46	4,39	4,08
	STD	1,00	0,68	0,81
Overall	**MOS**	**3,67**	**4,44**	**4,24**
	STD	**0,87**	**0,56**	**0,63**

Table 2: The evaluation results with regard to naturalness.

		Experiment 3 (paragraph-level)				
		Quality	Ease of listening	Pleasant-ness	Understand-ability	Pronunci-ation
Non-expert listeners	**MOS**	3,57	3,69	3,67	3,75	3,47
	STD	0,76	0,83	0,86	0,70	0,78
"Expert" listeners	**MOS**	3,54	3,64	3,53	3,72	3,48
	STD	0,84	0,87	0,84	0,75	0,83
Overall	**MOS**	**3,62**	**3,78**	**3,96**	**3,80**	**3,46**
	STD	**0,55**	**0,75**	**0,83**	**0,59**	**0,68**

Table 3: The evaluation results with regard to speech flow.

In order to investigate how well the resulted TTS system captures the specifics of the targeted domain, namely the news reports, we asked the subjects to rate the appropriateness of the system for different application areas, grading with 5 for perfect for the domain, and with 1 for inappropriate. The results are illustrated in the following table.

Application Domain	Appropriateness (1:poor 5:excellent)
News Portals	4.13
Telecom Applications	3.97
Accessibility Tools	4.5
Audio Books	3.76

Table 4: Mean Opinion Score for rating the appropriateness of the resulted TTS according to the application area.

The latter experiment provides a clear evident that the resulted TTS system captures efficiently the specifics of the targeted domain, providing at the same time enough adversity for coping with other similar domains.

Nevertheless, other domains, such as book reading, are more demanding areas with many aspects that could not be covered by the resulted speech database. It is also worth noting here that the high grade in the area of accessibility tools is mainly attributed to the fact that the resulted system delivers a high-quality synthetic speech and fulfils efficiently the most important requirements of the specific domain: intelligibility, robustness, consistency and pleasantness.

6. Discussion

In this paper we presented a methodology for designing and automatically producing an optimal corpus in the framework of a TTS system and for the specifics of the Bulgarian language. Our results depicted that the resulted TTS system, which incorporated the designed spoken corpus, performs significantly well, producing a high-quality near-natural synthetic speech. In the future we aim to investigate the differentiation of the TTS behavior with smaller databases, which would cover different aspects of our methodology, in order to identify possible prioritization in the aforementioned criteria and the level at which they affect the overall TTS performance.

Acknowledgements

The work presented in this paper has been co-financed by the European Regional Development Fund and by Greek national funds in the context of the INTERREG IIIA / PHARE CBC Programme 2000-2006 (an inter-regional cooperation programme between Greece and Bulgaria).
The authors would like to thank Prof. Elena Paskaleva, Ms. Irina Strikova and Ms. Aglika Ilieva Kroushovenska for their valuable help during the experiments, as well as the participants of the evaluation group for their useful feedback.

References

Alexander, R., & Zhobov, V., (Eds.), Revitalizing Bulgarian Dialectology, University of California Press/University of California International and Area Studies Digital Collection, Edited Volume #2, 2004.

Andersen O., Hoequist C., “Keeping Rare Events Rare”, Eurospeech 2003, vol. 2., pp. 1337-1340, 2003.

Balestri M., Pacchiotti A., Quazza S., Salza P., and Sandri S., “Choose the best to modify the least: a new generation concatenative synthesis system,” in Proceedings of the European Conference on Speech Communication and Technology (Budapest, Hungary), vol. 5, 1999, pp. 2291–2294

Black A., and Lenzo K., (2003) “Optimal Utterance Selection for Unit Selection Speech Synthesis Databases”, International Journal of Speech Technology, 6(4):357-363, October 2003, Kluwer Academic Publishers.

Bozkurt B., Ozturk O., Dutoit T., “Text Design for TTS Speech Corpus Building Using a Modified Greedy Selection”, Eurospeech 2003, pp. 277-280., 2003.

Franois H. and Boffard O., “The Greedy Algorithm and its Application to the Construction of a Continuous Speech Database”, in 3rd International Conference on Language Resources and Evaluation (LREC 2002), 2002, vol. 5, pp. 1420.1426.

Hauge, K. R., A Short Grammar of Contemporary Bulgarian, Slavica Publishers, Indiana University, USA, 1999

Hunt, A., and Black, A., “Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database”, in proc. Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Atlanta, pp. 373-376, 1996.

Iida, A., Campbell, N. (2001): "A database design for a concatenative speech synthesis system for the disabled", In SSW4-2001, paper 135.

Matousek J., Psutka J., Kruta J., “Design of Speech Corpus for Text-to-Speech Synthesis”, Eurospeech 2001, Alborg, 2001

Kishore S. P. and Black A., “Unit Size in Unit Selection Speech Synthesis”, Eurospeech 2003, pp. 1317-1320., 2003.

Lambert, T. (2006): "Automatic construction of a prosodically rich text corpus for speech synthesis systems", In SP-2006, paper 200.

Lewis E. and Tatham M., “Word and Syllable Concatenation in Text-to-Speech Synthesis”, Eurospeech 2001, vol. 2, pp. 615-618., 1999.

Malfrère, F., Dutoit, T., and Mertens, P., “Fully Automatic Prosody Generator For Text-to-Speech”, in proc. Intl. Conf. on Speech and Language Processing, pp. 1395-1398, 1998

Möbius B., “Corpus-Based Speech Synthesis: Methods and Challenges”, Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Univ. Stuttgart), AIMS 6 (4), pp. 87-116.,2000.

Möbius B., “Rare events and closed domains: Two delicate concepts in speech synthesis,” International Journal of Speech Technology, vol. 6, no. 1, pp. 57–71, 2003

Moulines, E., and Charpentier, F., “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones”, Speech Communication, 9: 453–467, 1990

Nagy A., Pesti P., Németh G., Bıhm T., “Design issues of a corpus-based speech synthesizer”, Hungarian J. Commun., 6: 18-24, 2005.

Raptis S., Tsiakoulis P., Chalamandaris A. and Karabetsos S., “High Quality Unit-Selection Speech Synthesis for Bulgarian”, In Proceedings of SPECOM2009, pp:388-393, St. Petersburg, 21-25 June 2009, Russia.

Schroeter, J., Basic Principles of Speech Synthesis, in Springer Handbook of Speech Processing, Benesty, J., Sondhi, M. M., and Huang, Y., (Eds.), Springer-Verlag, Berlin Heidelberg, 2008

Schweitzer A., Braunschweiler N., Klankert T., Möbius B., Sauberlich B., “Restricted Unlimited Domain Synthesis”, Eurospeech 2003, pp. 1321-1324., 2003.

Taylor, P. A., Analysis and synthesis of intonation using the tilt model, Journal of the Acoustical Society of America, 107, 4, 1697–1714 (2000)

Tokuda K., Yoshimura T., Masuko T., Kobayashi T., Kitamura T., Speech parameter generation algorithms for HMM-based speech synthesis, Proc. of ICASSP, pp.1315-1318, June 2000.

Yi, J.R.W., Glass, J.R., “Natural-Sounding Speech Synthesis using Variable-Length Units”, Proc. ICSLP-98, Sydney Australia, Vol. 4, pp. 1167-1170, 1998.