High Quality Unit-Selection Speech Synthesis for Bulgarian[1]

Spyros Raptis, Pirros Tsiakoulis, Aimilios Chalamandaris and Sotiris Karabetsos

Voice and Sound Technology Department
Institute for Language and Speech Processing / R. C. "Athena", Athens, Greece
{spy, ptsiak, achalam, sotoskar}@ilsp.gr

Abstract

We present the design, implementation and assessment of a high-quality unit-selection speech synthesis system for Bulgarian which, to the best of the authors' knowledge, is the first such synthesizer for the Bulgarian language. The paper presents implementation details for each component of the system, emphasizing on language-specific aspects and arguing that corpus-based unit-selection synthesis offers a much better framework for dealing with the phonetic and prosodic characteristics of the language than rule-based or fixed-inventory concatenative methods. The paper concludes with the presentation of the evaluation setting and results which provide clear evidence of the quality level achieved.

1. Introduction

Text-to-speech synthesis systems convert textual input into synthetic voice signals. Offering the promise of natural and intuitive human-computer interaction, 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. Today's speech synthesizers are widely employed in the context of assistive aids and tools, telecommunication systems, entertainment, and education.

Most often, speech synthesizers are considered to be composed of two main parts [1]: a frontend and a backend. The frontend is, in effect, a text-processing component working on a symbolic level and converting the input text into control information. Its main task is to perform text normalization and linguistic analysis in order to produce phone-based information and prosodic annotation. The backend is mainly a signal-processing component that uses the control information to generate the synthetic speech. Depending on the application, frontend processing may also need to include additional tasks, such as detection of document structure, interpretation of text markup etc. Similarly, backend processing may also be extended to include tasks such as voice post-filtering to simulate various effects like whispering, echo etc.

From the early days of the barely intelligible, mechanically-sounding speaking machines, up to the present day of near-natural speech synthesizers, a wide range of approaches have been proposed for the implementation of the synthesis backend. These are roughly divided into two broad categories (even though they often share technological background and algorithms), namely rule-based systems employing rules to drive models of speech production and corpus-based (or data-driven) systems that generate speech by manipulating and concatenating stored segments of pre-recorded human speech. Most of the today's top quality speech synthesizers fall into the corpus-based category; they employ sophisticated unit-selection techniques in order to select the optimal segments among thousands of instances from a large repository of pre-recorded speech. The selected segments are then concatenated using standard signal processing techniques (usually pitch-synchronous overlap-add) to generate the synthetic speech.

Besides some earlier works that have led to some first text-to-speech systems of limited quality for Bulgarian, the most recent advance is the SpeechLab 2.0 system [2]. It places special emphasis on the text processing component and employs a rule-based approach implemented by a pipe of finite state devices. Text processing takes place in three phases. At the first phase, part-of-speech annotation is performed. The second phase applies an accentual dictionary and the third phase proceeds with 98 contextual rules for phonetization, accent determination and unknown words processing, concluding with a set of 91 rules for prosody annotation. At its backend, the system uses the FD-PSOLA method [3] to perform diphone manipulation and concatenation. Nevertheless, despite the thorough text processing that is performed and the extensive set of rules employed, the resulting speech still has a "robotic" timbre and offers rather limited naturalness. This can be attributed to the known limitations of the backend processor used: a limited inventory of speech units, degradation due to significant signal processing required for pitch modification, phase and spectral mismatches and so on.

The synthesizer presented in this paper follows the unit-selection paradigm. Employing a corpus-based rather than a rule-based approach, the system is able to handle more efficiently a number of critical issues relating to phonetization and prosody, since it does not need to rely on explicitly devised rule sets but on actual instantiations of speech as present in the speech database. By including a much more extensive database of speech units and by implementing efficient unit-selection mechanisms for picking the optimal ones, it manages to avoid the shortcomings of typical limited inventory systems. The system is able to produce near-natural synthetic speech that can fully satisfy the requirements of most real-life applications. It achieves remarkable naturalness and high intelligibility, as also demonstrated during the evaluation. To the best of the authors' knowledge, this is the first unit-selection text-to-speech system presented for the Bulgarian language.

The rest of the paper is organized as follows. The following section gives an overview of the system, along with a discussion on some aspects of Bulgarian language and the advantages of the corpus-based unit-selection approach in

[1] Samples of synthesized utterances are available online at: http://speech.ilsp.gr/specom2009.

dealing with them. Section 3 provides some more detailed information on the different modules of the system, offering further implementation details. Section 4 describes the evaluation process and the results obtained. Finally, the last section presents conclusions and closing remarks.

2. Overview of the system

An overview of the synthesis system is given in Figure 1. The frontend component follows the typical architecture of most synthesizers [1], including modules to handle text normalization and phonetic analysis. The backend component employs the unit-selection concatenative speech synthesis paradigm employing a large annotated database of pre-recorded natural speech. More information is provided in the following paragraphs. The presentation is mainly concerned with language-specific issues, avoiding any details for other modules typically found in speech synthesizers.

2.1. Frontend Processing

2.1.1. Text Processing

The text processing module is responsible for tokenizing, normalizing and, when required, tagging the input text.

Tokenization refers to the identification of word and sentence boundaries. As in many other languages, the main problem in sentence boundary detection in Bulgarian is the ambiguity of the punctuation marks in their different functions, especially the full stop marked by a point. A point could serve many different purposes in a string, including decimal separator in numbers, a letter separator in abbreviations and acronyms and so on. However, a point mark is usually not used as a rank separator in numbers in Bulgarian. The system presented here uses a set of heuristic rules to identify sentence boundaries based on the neighboring elements in the string and their characteristics.

Text normalization refers to appropriately handling special cases of tokens in the input text, such as acronyms, abbreviations and numerals [4].

Regarding numerals, any group of digits could form a numeral. Bulgarian typographic standards require a space between every third digit and a decimal comma. However, a comma is also often used to separate 3-digit groups. A comma is also used to denote decimals digits, not a point.

In general, the numeral expansion is similar to English but it is not identical. In plural numbers ending in a form of един (one), agreement is usually plural, by sense. In mathematical usage and when counting without counting any specific object the neuter forms are used.

A dot is used to separate hours from minutes. The noun часа (hours) may be abbreviated into ч. In dates, ordinals are used for the day of the month. The year is expressed by an ordinal determining the noun година (years), usually abbreviated as г.

Except as an end-of-sentence indicator, the period mark is further used in abbreviations where the first letter(s) of the word is kept and the rest is truncated. In abbreviations where the middle of the word is truncated, with a hyphen replacing the truncated part, no period is used. However, the monetary unit лева (levs) is abbreviated лв., with a period, and metric units of measure are abbreviated without a period. Acronyms are formed by capitalizing the first letters and usually contain no period marks.

The system is able to identify and perform basic handling of numerals (cardinals, ordinals, decimals and fractions) as well as date and time strings. In order to handle acronyms and abbreviations, the system incorporates a list of the most frequent such tokens encountered in Bulgarian texts.

2.1.2. Phonetization

A phonetic transcription module is an indispensable component at the heart of any speech synthesis engine. The role of this highly language-dependent module is to map text input into the corresponding sound patterns of the language.

Modern Bulgarian is composed of 6 vowels and 24 consonants. A set of 41 phonemes can be considered as sufficient to describe the sound system of the language at the phonological level (apart from composite sounds such as /ts/, /tʃ/ and /dʒ/).

A rather limited set of rules can be formulated to obtain a broad phonetic transcription of the Bulgarian language. However, most speech synthesis methodologies, especially the rule-based ones, strongly depend on the precision of the transcription. For example, not only allophonic variations, but also more complex phenomena in the realization of speech, such as coalescence, reduction and palatalization need to be registered with enough precision by the phonetic transcription

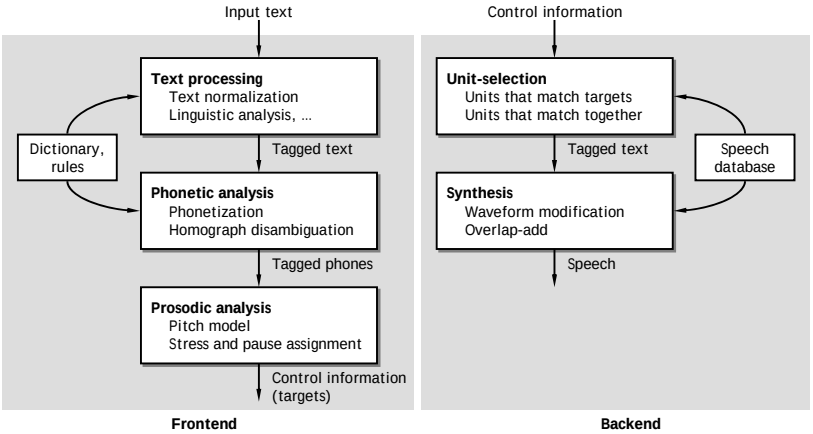

***Figure 1**: Overview of the synthesis system (adapted from [1]).*

module in order to be handled correctly by the backend of the synthesis system.

Bulgarian's six vowels tend to merge as, in an unstressed position the open vowels are raised towards the corresponding close vowels. However, the coalescence is of variable extent and not always complete depending on various factors [5]. A clearer distinction tends to be maintained in the syllable immediately preceding the stressed one. In the literary language, it is considered substandard for the front vowel /ɛ/, while reduction of the central and back vowels /ɑ/ and /ɤ/ is quite widespread [4]. The merger of the pair /ɑ/ - /ɤ/ may also be considered as common while the reduction of /ɔ/ is not usual for the literary norm and is regarded as dialectical. A coalescence of /ɛ/ and /i/ is not allowed in formal speech whereas unstressed /ɛ/ is both raised and centralized, approaching /ɤ/. Vowels also tend to become nasalized when followed by the combination nasal consonant + fricative.

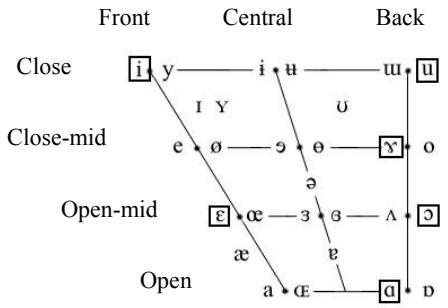

***Figure 2**: The Bulgarian vowel chart (adapted from [9]).*

In Bulgarian, palatalized consonants are considered as separate phonemes. However, they may, in some cases, be positionally conditioned. К /k/, г /g/ and х /x/ tend to be palatalized before /i/ and /ɛ/, and the realization of the phoneme л /l/ varies along the same principles: one of its allophones occurs in all positions, except before the vowels /i/ and /ɛ/, where a more "clear" version with a slight raising of the middle part of the tongue occurs.

The strong dependence of the overall performance of most speech synthesis approaches on the accuracy of the phonetic transcription significantly raises the expectations from the transcription module, necessitating the use of a narrow phonetic transcription. However, any attempt to obtain a complete and coherent set of rules for narrow phonetic transcription is confronted with the complex and delicate problem of determining the context of such variations, as shortly discussed above. Furthermore, such variations are not mere switches; an open vowel's tendency to reduce to its corresponding close vowel is manifested in different ways and variable extent depending on a set of parameters [5]. Due to their inherent difficulty, these questions do not always find widely accepted answers.

Furthermore, dialect features (such as the coalescence of /ɛ/ and /i/ [6]), speaker-dependent variations in pronunciation (such as the palatalization of ж /ʒ/, ш /ʃ/, ч /ʧ/ and дж /ʤ/), and phenomena related to speaking style (such as the realization of velarized allophone of /l/ in many young people's speech), add further to the complexity.

Nevertheless, the difficulties encountered are not only the ones stemming from the language itself. Variations among different speakers and speaking styles are an additional source of concern. One should notice that, when developing a particular synthetic voice, the system would only need to capture and imitate the specifics of that single speaker. While this may appear to relieve some of the burden of covering the entire range of possibilities, it actual implies that various system components should be tailored (or, at least, fine-tuned) to the particular speaker's characteristics. But that, of course, is rarely the case.

The above provide clear indications of the constraints of rule-based approaches and the advantages of corpus-based ones.

Contextual variations, extent and speaker-dependency are implicitly addressed by unit-selection synthesizers. A broad phonetic transcription, coupled with appropriately tuned join weights on context in the unit selection formula, can effectively mimic speech variability, provided, of course, that the underlying speech database offers sufficient coverage of such phenomena and a necessary degree of redundancy.

The presented synthesis system internally uses broad phonetic transcription. A pronunciation dictionary is used to retrieve the phonemic form (and the stress position) for each word in the input. This is complemented by a compact set of rules to handle out-of-vocabulary words. The limitations of broad phonetic transcription in capturing the variability in speech units is effectively compensated by the phonetically and contextually balanced speech database employed by the system, which ensures that instances of speech units in various contexts are sufficient represented.

2.1.3. Stress

Bulgarian word stress is dynamic and lexical (rather than fixed), and it is not usually signified in written text (with few notable exceptions used for disambiguation).

One-syllable words are stressed on their unique vowel, with the exception of functional words such as clitics. Clitics, which may form clusters, do not carry stress but are coupled to a neighboring stressed word.

Free word stress requires that the phonetization module can properly identify the stress location for each word. As is often the case, the synthesis system presented in this paper relies on lexica for registering the stress position of the words in the input string, along with their pronunciation.

In Bulgarian, stress is also distinctive. It is used to discriminate between different meanings in the case of homographs. However, such a disambiguation is particularly hard to accomplish automatically, since the alternatives may be of the same part-of-speech and share the same grammatical forms. Thus, the disambiguation would need to take place at a higher level of analysis. A common approach in dealing with homographs, which is also the one employed by the current system, is to either add a (possibly weaker) stress at both candidate positions or to completely omit it and rely on the listener to resolve the ambiguity.

2.1.4. *Prosody modeling*

Explicit pitch models of prosody, such as the Fujisaki model [10], the Tilt model [10] or other models generating stylized intonation patterns, are commonly used in speech synthesis. These models, however, present significant drawbacks as they fail to sufficiently capture and represent the variability and richness of the patterns found in human speech, no matter how many refinements and adjustments are made to the models.

The very problem of designing a generic, speaker-neutral model of intonation is ill-formulated as there is no such thing as a single "correct" intonation for a given sentence (even for a single speaker) and there is certainly no such thing as a "reference" intonation model able to capture all the speakers of a given language. The fine-structure of intonation patterns, which are closely linked to the naturalness perceived in human speech, is highly speaker-dependent. Furthermore, in the context of a text-to-speech system the intonation model does not operate in isolation. The pitch contours it produces are combined with segmental information (or, in corpus-based systems, with units from the speech database) to produce the output speech.

The above facts provide clear indication that, in corpus-based systems, intonation and the units in the speech database cannot be treated separately. This fact is recognized in data-driven intonation models [7] which resort to the speech database not only to retrieve speech units but also to acquire actual pitch patterns of the specific speaker.

This is the approach adopted in the presented synthesis system, where intonation modeling has been integrated in the unit selection process. Apart from its other characteristics, a set of prosodic features is also assigned to each unit in the database that describe the unit's pitch level and position in relation to stresses, pauses or the end of the sentence. At run time, these features are also taken into account in the unit selection stage along with the spectral features, and contribute to the target cost (from the desired features) and the join cost (a measure of how similar two units are at their edges), as explained in the respective section.

2.2. Backend Processing

2.2.1. *The Speech Database*

The current state-of-the-art unit-selection synthesizers produce highly intelligible, near natural synthetic speech. However, this usually comes at the cost of large resource repositories and increased processing power. This is because unit selection speech synthesis relies on large speech databases. The larger the database is the more natural the synthetic speech is. The speech database usually consists of naturally spoken utterances, carefully annotated to the unit level. Each utterance comes from a text-corpus designed to cover as many units as possible in different phonetic and prosodic contexts. The resulting repository of speech units may have little or great redundancy, on which speech variability and overall quality some times depend.

The corpus design for a corpus-driven text-to-speech system is one of the most important tasks, since the overall quality of the final system depends on the principal data it uses.

The main objective of the approach employed for the corpus design task was to achieve sufficient coverage of the significant language phenomena identified. This can be projected to the following set of complementary goals: (i) phonetic coverage, (ii) prosodic coverage, and (iii) controlled redundancy.

Phonetic coverage refers to the inclusion of all possible principal speech units (in our case, diphones) that are encountered in the spoken language, including a set of diphones necessary for the utterance of foreign words. The goal of *prosodic coverage* is to cover as many different prosodic environments as possible, as they are distinctively defined by the prosody modeling component of the system. This means that, given the prosody generation engine and all the possible prosodic phenomena that it can identify and model, the resulted corpus must cover as many as possible different environments where the aforementioned phenomena occur. The *controlled redundancy* goal refers to ensuring that the previous goals do not lead to an exceedingly large corpus with unnecessary redundancy. All the above goals are properly coordinated to strike the necessary balance between corpus size and redundancy.

Based on the above, from an initially collected corpus of about 46 million words, a subset of 600 sentences was selected containing about 5K unique words and producing a spoken corpus of 60K diphones instances.

2.2.2. *Unit selection*

The unit-selection module [8] is considered to be one of the most important components in a corpus-based unit-selection synthesis system. It provides a mechanism to automatically and efficiently select the optimal sequence of units that participate in the production of the final speech output.

The criterion of optimality is the minimization of a total cost function which is defined by two partial cost functions, namely, the target cost and the concatenation (join) cost function. The target cost function measures the similarity of an applicant unit with its predicted specifications and the concatenation cost function accounts for the acoustic matching between pairs of consecutive candidate units. For data-driven prosody modeling, as is the case in the presented system, prosodic considerations are integrated into the unit-selection process and the two costs (target and join) also take into account not only spectral criteria and absolute pitch differences, but also the prosodic features of the units in the database. Selecting the optimal sequence usually employs a thorough search (a Viterbi search) and comparison, through calculations of similarity measures, between all available units.

2.2.3. *Signal processing*

After the optimal units have been selected from the speech database, the system employs standard time-domain pitch-synchronous overlap-add (TD-PSOLA) techniques to concatenate them and generate the output signal [3].

This method roughly consists of extracting the pitch periods of a voiced speech signal, windowing each segment with a Hanning window centered on every glottal closure point and then moving the segments closer or further apart to achieve pitch lifting or lowering. The manipulated segments of the consecutive units are overlapped and added to perform

		Experiment 1 (sentence-level)			Experiment 3 (paragraph-level)				
		Naturalness	Ease of listening	Articulation	Quality	Ease of listening	Pleasant-ness	Understand-ability	Pronunci-ation
Non-expert listeners	**MOS**	3,53	4,41	4,13	3,57	3,69	3,67	3,75	3,47
	STD	0,96	0,66	0,77	0,76	0,83	0,86	0,70	0,78
"Expert" listeners	**MOS**	3,46	4,39	4,08	3,54	3,64	3,53	3,72	3,48
	STD	1,00	0,68	0,81	0,84	0,87	0,84	0,75	0,83
Overall	**MOS**	**3,67**	**4,44**	**4,24**	**3,62**	**3,78**	**3,96**	**3,80**	**3,46**
	STD	**0,87**	**0,56**	**0,63**	**0,55**	**0,75**	**0,83**	**0,59**	**0,68**

***Table 2**: The evaluation results with regard to naturalness (experiment 1) and speech flow (experiment 3).*

their concatenation. No special smoothing or post-processing is employed in unit join boundaries.

3. 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.

A group of 30 native Bulgarian speakers participated in the evaluation. Among them, 10 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 experiments were performed in an unsupervised setting, after the necessary guidelines and instructions have been provided to the participants. They were able listen to each stimulus more than once. The majority of the participants were able to complete the entire set of experiments within about 1 hour.

Experiment 1: Naturalness. The aim of the first experiment was to evaluate the performance of the TtS system in terms of naturalness. The Mean Opinion Score (MOS) was used as the subjective scoring method. The stimuli consisted of 35 randomly selected, medium-sized sentences with an average of 13 words per sentence. The sentences where synthesized using the text-to-speech system, and the listeners were asked to rate three quality dimensions for each sentence by scoring on a scale of 1 to 5 for each dimension. To ensure consistency in responses, each grade was assigned a label as shown in Table 1 below:

	Naturalness. *How close did you perceive the synthetic speech to be to natural (human) speech.*	**Ease of listening**. *How difficult was it to follow what was being said.*	**Articulation**. *Is the utterances well articulated or are there any irregularities.*
1	Unnatural	No meaning understood	No
2	Inadequately natural	Effort required	Not very clear
3	Adequately natural	Moderate effort	Fairly clear
4	Near natural	No appreciable effort required	Clear enough
5	Natural	No effort required	Very clear

***Table 1**: Labelling of the scale for each aspect under evaluation.*

Table 2 summarizes the mean scores (MOS) and the standard deviations of the responses, discriminating between "expert" and "non-expert" listeners. Interestingly enough, the opinions of the two groups where highly consistent.

It is worth noting that the "ease of listening" and the "articulation" received remarkably high grades, which were consistent among both experts and non-experts. Furthermore, the overall score for "naturalness" which lies near 4 is particularly high, considering that 4 corresponded to "near natural".

Experiment 2: Intelligibility. The aim of this phonetic task was to evaluate the synthesis system in terms of intelligibility. The Diagnostic Rhyme Test (DRT) was employed, which provides a widely used index for diagnostic and comparative evaluation of the intelligibility of single initial or final consonants.

The stimuli consisted of 33 groups of two or three words each, some of which were nonsense. The words in each group were only differentiated in one letter. For each group, the participants were presented with the list of words and one of them was synthesized and played back. They were then asked to select which word from the list they listened.

Some examples of such word pairs are: виц/вид/виж, печен/сечен/речен, росата/косата etc.

In the vast majority of cases, all participants were able to correctly match the stimulus with the respective word in the list.

Experiment 3: Speech Flow. The aim was to evaluate the TtS system in terms of speech flow. The stimuli consisted of 5 randomly selected paragraphs with an average of 6 sentences (or 83 words) per paragraph.

Various aspects of the synthetic speech were addressed for each paragraph: smoothness, naturalness, pleasantness, clarity, and appropriateness of the synthetic speech. The target was to obtain feedback on the overall listening experience as perceived at a level higher than a single sentence.

The paragraphs where synthesized using the text-to-speech system and the listeners were asked to rate the above quality dimensions for each paragraph by assigning a score on a scale of 1 to 5 for each dimension. To ensure consistency in responses, each grade was again assigned appropriate labels. The MOS was utilized as the assessment criterion. The results are summarized in Table 2. Again, no appreciable deviation

was observed between the responses of the expert and non-expert groups.

Questionnaire: Application Areas. A final set of questions was used to capture the participants' opinion regarding how appropriate it would be for the synthesis system to be used in different application areas.

The participants were given the option not to answer to any of those questions if they felt that they could not offer an informed response. The ones that did answer, used a rating scale of 1 to 5 with 1 corresponding to "inappropriate" and 5 to "completely appropriate".

The results obtained for the different application areas are summarized in the figure below ("Bad/Poor" corresponds to rates 1-2, "Fair" corresponds to rate 3, and "Good/Excellent" corresponds to rates 4-5).

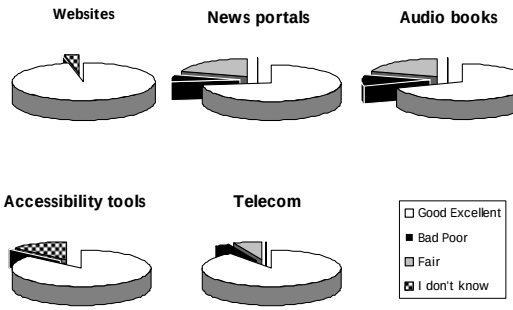

Figure 3*: Responses regarding the suitability of the synthesis system in different application areas.*

The results offer a clear indication that the synthesis system is highly regarded as a very appropriate tool in almost all the application areas. It is worth noting that the lowest score was received for "Audio books". This was expected since book reading not only represents one of the most demanding areas for text-to-speech technology, but also because the formal speaking style usually employed in synthetic voices is often less appropriate.

4. Conclusions

Unit-selection methods are now considered to offer the most efficient framework for the implementation of high-quality speech synthesizers.

In this paper we presented the design, implementation and assessment of a high-quality unit-selection speech synthesis system for Bulgarian. The system exploits data-driven techniques to bypass the problems that rule-based methods encounter when faced with inherently difficult language-dependent tasks such as phonetization and prosodic modeling. Data-driven techniques alleviate the need for making unnecessary assumptions and for explicitly formulating rigorous rule sets to capture complex language phenomena. Instead, they direct the system to a real spoken corpus, the speech database, to draw actual speech instances and rich prosodic patterns.

The system manages to produce synthetic speech of very high quality, achieving high degrees of naturalness and intelligibility, as also confirmed through a range of acoustic experiments performed. The experiments offered clear indication that the synthesis system is considered as a very appropriate tool for almost any application areas.

5. 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 the valuable help they offered during the project, as well as the participants of the evaluation group for their useful feedback.

6. References

[1] 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

[2] Andreeva, M., Marinov, I., and Mihov, S., "SpeechLab 2.0 -- A High-Quality Text-to-Speech System for Bulgarian", *Proceedings of the Intl. Conf. on Recent Advances in Natural Language Processing (RANLP2005)*, pp. 52-58, Borovets, Bulgaria, September 2005.

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

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

[5] Anderson, J. M., "The representation of vowel reduction: non-specification and reduction in Old English and Bulgarian", *Studia Linguistica*, 50. 91-105, 1996

[6] 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. (http://repositories.cdlib.org/uciaspubs/editedvolumes/2)

[7] 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

[8] 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.

[9] Maddieson, I., *Patterns of Sounds*. Cambridge: Cambridge University Press, 1984.

[10] Fujisaki, H., and Kawai, H., "Modeling the dynamic characteristics of voice fundamental frequency with applications to analysis and synthesis of intonation", in *Working Group on Intonation, 13th International Congress of Linguists* (1982).

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