The ILSP Text-to-Speech System for the Blizzard Challenge 2012

Spyros Raptis[1,2], Aimilios Chalamandaris[1,2], Pirros Tsiakoulis[1,2], Sotiris Karabetsos[1,2]

[1] Institute for Language and Speech Processing / Research Center "Athena", Athens, Greece
[2] INNOETICS LTD, Athens, Greece

{spy,achalam,ptsiak,sotoskar}@ilsp.gr

Abstract

This paper describes ILSP and INNOETICS Speech Synthesis System entry for the Blizzard Challenge 2012. A description of the underlying system and techniques used are provided, as well as information about the voice building process and discussion on the obtained evaluation results. Additional focus will be given to new processes or techniques we used this year in comparison to our previous participations, and we will also discuss the results of the new section of the Blizzard Challenge which aims to investigate the abilities of the participating TTS systems to cope with audio books and expressive speech synthesis.

Index Terms: speech synthesis, unit selection, speech evaluation, Blizzard Challenge 2012, audio books, librivox, expressive speech synthesis.

1. Introduction

This is the third participation of the Speech Synthesis Group of the Institute for Language and Speech Processing (ILSP), Athens, GREECE, and INNOETICS LTD to the Blizzard Challenge. This paper presents the system used for the ILSP/INNOETICS entry to the Blizzard Challenge 2012 competition.

ILSP has been in the state-of-the art in text-to-speech research in Greece for almost two decades, having developed TtS engines for the Greek language based on all the major approaches: formant rule-based (e.g. [1]), diphone (e.g. [2]), and unit-selection. Recently, the Speech Synthesis Group at ILSP has developed the first TtS prototype for Greek employing statistical/parametric speech synthesis with HMMs [3].

The system entry for the Blizzard Challenge 2012 competition is based on the core TtS engine by ILSP, as enhanced with speech tools and techniques by INNOETICS Ltd, a spin-off company offering commercial solutions based on the core technology. The initial design of the engine has been initially carried out based on the Greek language. However, as a corpus-based system, most modules are language-independent, with already successful migrations and customizations to other languages such as Bulgarian, offering equally high-quality results [4]. A scaled-down, low-footprint version of this system has also been developed for mobile environments [5].

This paper is organized as follows. First, we describe the system with some detail, focusing on specific modules. In section 3 we describe the voice building process and specific adaptations that were necessary for this challenge, while in sections 4 and 5 we present the results and we discuss them respectively.

2. System Overview

Although the architecture of our TTS system is given in previous publications, for the sake of completeness we present it here as well.

Our TtS System follows a typical concatenative, unit-selection architecture as depicted in Figure 1.

The two main modules incorporated by the system are the Natural Language Processing (NLP) and the Digital Signal Processing (DSP) component.

2.1. The NLP Subsystem

The NLP component is mainly responsible for parsing, analyzing and transforming the input text into an intermediate symbolic format, appropriate to feed the DSP component. Furthermore, it provides all the essential information regarding prosody. It is composed of a word- and sentence- tokenization module, a text normalizer, a letter-to-sound module and a prosody generator.

All these subcomponents are necessary for the disambiguation and proper expansion of all abbreviations, numerals and acronyms, for the correct word pronunciation, and also for the detection and application of the rich set of distinctive features of the speech signal, closely related to prosody.

2.1.1. Tokenization

The input text is fed into the *parsing module*, where sentence boundaries are identified and extracted. This step is important since all remaining modules perform only sentence-level processing.
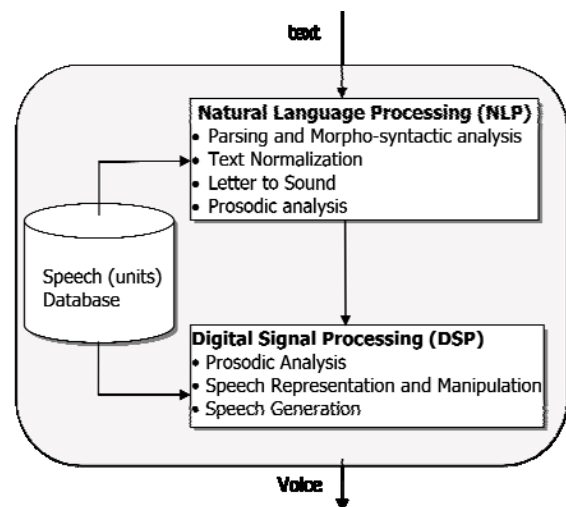

Figure 1: *Overall system architecture.*

2.1.2. Text normalization

The identified sentences are then fully expanded by the *text normalization* module, taking care of numbers, abbreviations and acronyms.

2.1.3. Letter-to-sound conversion

The *letter-to-sound* module transforms the expanded text in an intermediate symbolic form related to phonetic description. For English we used a lexicon-based approach complemented by a set of automatically-derived rules to handle out-of-vocabulary words. The rules were extracted using a method similar to the one described in [6]. An exception dictionary was also included. This was our first US-English accented voice and therefore special customization of the letter-to-sound module had to be performed during this year's challenge.

2.1.4. Prosody prediction/specification

The overall approach used for handling prosody in this version of the system, is a stripped subset of the one used for the Greek version of the system. No particular customization has been performed for the English language, except from some minor adaptations to take into account the secondary stress which seems to be much more important in English than it is in Greek.

No explicit prosodic modeling is performed, in terms of target pitch values or duration models. The approach employed for prosody is taking into account the distance of a diphone from prosodically salient units in its vicinity such as stressed syllables, pauses, and sentence boundaries, and the type of these units discriminating between declarative, interrogative and exclamatory sentences. This information is fed to the target cost component of the overall cost function in the unit-selection module. The main motivation behind such a rather plain approach is that naturalistic prosody patterns can be expected to emerge by the corpus through the unit selection process, assuming that the corpus is large enough and that the major factors affecting prosody have been taken into account.

There was no explicit bias in our system towards the selection of consecutive database units at the syllable or any other level, other than the implicit favoring of consecutive units by the unit-selection procedure due to their low join cost.

2.2. The Acoustic Subsystem

The DSP component comprises of the unit selection module and the signal manipulation module. The ILSP TtS system relies on a Time Domain Overlap Add method for speech manipulation. The DSP component also includes the unit selection module, which performs the selection of the speech units from the speech database using explicit matching criteria. More details about each of these modules are given below.

2.2.1. Unit-selection

The unit selection module is considered to be one of the most important components in a corpus-based unit selection concatenative speech synthesis system. It provides a mechanism to automatically select the optimal sequence of database units that produce the final speech output, the quality of which depends on its efficiency. The criterion for optimizing is the minimization of a total cost function which is defined by two partial cost functions, namely the target cost and the concatenation cost function [7].

More specifically:

- *the target cost components*: two target cost components are used: one that accounts for the similarity of the phonetic context (spanning 2 phones on each side) and one that accounts for the similarity of the prosodic context, the latter being formulated as described in section 2.1.4 above.

- *the join cost components*: two join cost components are used: one that accounts for pitch continuity and one that accounts for spectral similarity. While the system currently employs Euclidean distance on MFCCs, there is ongoing research in the group to move to spectral join cost calculation based on one-class classification approaches [8].

The weights for each component of the cost function are manually tuned and are phoneme dependent.

2.2.2. Pitch-smoothing

After the candidate units have been selected from the speech database, only minor modification is performed to the resulting pitch contour in order to remove any significant discontinuities at the boundaries of consecutive voiced units and to smoothen the overall pitch curve. A polynomial interpolating function (similar to low-pass filtering) is used on the pitch contour to perform the smoothing.

2.2.3. Waveform generation and manipulation

A custom Time Domain Overlap Add (TD-OLA) method is used to concatenate the selected and apply the smooth pitch contour, in a pitch synchronous method.

2.2.4. Hybrid approach for unit selection

This year, we decided to further investigate a hybrid approach for the unit selection module. As a first attempt to this approach we developed an external tool that provided the final unit selection path by taking into account aside from our unit selection criteria and weights, a set of questions derived from the corresponding HTS system we have integrated [3]. Since the initial results were not significantly better and due to time constraints, we decided not to adopt this approach for our participation in this challenge mainly because we wanted to investigate specific aspects of our already developed concatenative TTS system and in order to avoid possible errors and bugs that could have been introduced during this fast process. However, this approach is something we are going to further investigate in the immediate future.

3. Building a voice from the LibriVox audio data

The following paragraphs describe the process of building the Blizzard 2012 voices for use with ILSP's TTS system. The US-English voice for the Blizzard 2012 challenge was built using the provided audio data. This data was provided to the Blizzard participants by TOSHIBA and it includes the segmented audio files of three different audio books, as they were narrated by the same voice talent.

3.1. Audio Preprocessing

The first step was the amplitude normalization of the audio files in order to alleviate large amplitude mismatches during synthesis. For the creation of the database we used the provided audio data sampled at 16 KHz, together with their corresponding transcription.

Since the provided audio data was a result of different recordings with different equalization settings, and with possibly different hardware, we decided to equalize the audio recordings by using spectral equalization techniques in order to achieve the same average spectral content for every utterance. This method, although it provided similar sounding between the audio recordings at first, during synthesis we noticed that spectral discontinuities were obvious between segments from different audio books, and therefore we decided to limit our TTS system to use audio data from within a single audio book for every sentence it would synthesize.

3.2. Building the Voices

This section provides a description of the steps we followed to build the Blizzard Challenge 2012 voice. As mentioned above, we built a different database for every audio book provided. This decision was made in order to investigate whether the performance of our system would be affected by the rather inconsistent settings of recordings between different books, or even between chapters of the same book. The different databases could be combined or function independently.

3.2.1. Labeling

For the phonetic annotation of the speech corpus, we used the data provided by TOSHIBA. We therefore decided to use the provided phoneme set as well as the provided phonetic transcription for the production of the stimuli. As far as the prosodic annotation is concerned, we used our own custom label set which takes into account the punctuation, stress, intonation and the phonetic attributes of every phoneme. It is entirely language independent since it is based on the phonetic level of the language, and it can be used for every language without any modification.

3.2.2. Segmentation

For the segmentation of the audio data we used annotation data provided by TOSHIBA, and no automatic or manual correction was performed, nor did we normalize the silences or pauses within the utterances. However, in order to minimize the errors from the segmentation process, we decided to remove from our database all the sentences of which the final recognition score was less than 100%.

Table 1. *The audio data length for each audio book and the resulted database after the pruning mentioned.*

	# Waves	Length Hours	# Waves 100% ASR	Further Prunning	DB Length Hours
Book #1	7499	15,2	5030	10%	8,6
Book #2	7020	13,0	4943	29%	6,0
Book #3	5214	6,3	4067	29%	3,3
Book #4	7609	11,6	5089	30%	5,1

3.2.3. Pruning

Due to time limitations and the size of data, only automatic database pruning was performed based on two criteria: a) the recognition score of every sentence and b) average spectral content of every utterance. Any sentence that aligned with a score lower than 100% was removed, and any sentence the average spectral content of which was substantially different from the rest of the utterances was also removed. The latter

was performed by a k-means classification of the spectral contour of every wave file, which resulted in identifying with significant effectiveness the wave files that included imitation or role playing by the narrator. The above process resulted in an additional pruning of about 25% of the already pruned audio data (based on the 100% ASR rate criterion as mentioned above).

3.2.4. Pitch-marking

For pitch marking, we utilized the method we have developed and which is described in [10].

4. Evaluation Results

During Blizzard Challenge 2012 several aspects were put into evaluation with significant differences from previous challenges. The two main differences were the audio data provided, which was derived from expressively narrated audio books, and the new section of questions for investigating the performance of the synthesized speech in new fields such as emotion and intonation when coping with different types of text, such as book paragraphs. In total four different aspects were tested: a) naturalness, b) similarity to the original speaker, c) word error rate and d) appropriateness for audio books. The latter includes several different requirements such a system would need to meet, such as the level of emotion expressed by the system, the listening effort and other more detailed aspects such intonation, stress and silences manipulation.

In the following results our system is identified with the letter "I".

4.1. The Stimuli

As mentioned before, we created different audio databases for our TTS system, one for each audio book we had in our disposal, and although the wave files were spectrally equalized, we did not allow our system to use audio segments from different audio books for the creation of a stimulus, in order to avoid noticeable spectral discontinuities between segments. In order to do so, we tweaked our TTS system and for every stimulus, our TTS used the database with which the final unit selection score would be minimum, compared to the other databases. In practice this means that our system synthesized every stimulus 4 times (as many as the audiobooks) and it picked the one with the lowest overall cost in the unit selection process. Even though this means it used extra time and computational load, it did not really matter since the synthesis of the stimuli took place offline.

4.1.1. Naturalness

As far as the naturalness is concerned our system ranked at the 3rd and 2nd position, depending on the listeners group, achieving an average MOS of 3.3, a result which is similar to our previous participation in Blizzard 2011 [11] [12], where the audio data was designed and recorded for use with TTS and not as expressively as this year's audio data. This could actually be explained as ceiling effect with the natural audio data in the stimuli set, or as an actually very interesting result hinting that less supervised audio data can produce equally natural-sounding TTS systems with audio material that was especially designed and recorded.

Table 2 below, shows the Mean MOS-naturalness scores for this task, with additional breakdown information for the listeners groups. For all the displayed results in the following tables, system A denotes natural speech.

A more analytical look on the results leads us to an additional conclusion about our TTS system which behaves rather better when the context of application is similar to the domain of the training data. This can be depicted in Figure 2 where in stimuli from novels our system ranked 2nd with an average MOS 3.5 while in the same experiment with stimuli describing news our system ranked 3rd with an average MOS of 3.2.

Table 2. *Mean MOS-naturalness scores for Blizzard Challenge 2012 for all participating systems.*

	All Listeners	Paid Listeners	Online Volunteers	Speech Experts
A	4,7	4,7	4,7	4,7
B	3	3	3,4	3,5
C	3,8	3,8	3,6	3,9
D	2,2	2	2,7	2,8
E	1,6	1,5	1,8	1,7
F	3,4	3,3	3	3,1
G	2,5	2,4	2,7	2,9
H	2,6	2,5	2,8	2,5
I	3,3	3,3	3	3,2
J	1,9	1,6	2,3	2,6
K	1,6	1,6	1,8	1,9

In Figure 1 below, one can view the standard boxplots for the Mean opinion scores for naturalness for Blizzard 2012 (all listeners) while in Figure 2 the same results but for the Novel Stimuli only are depicted.

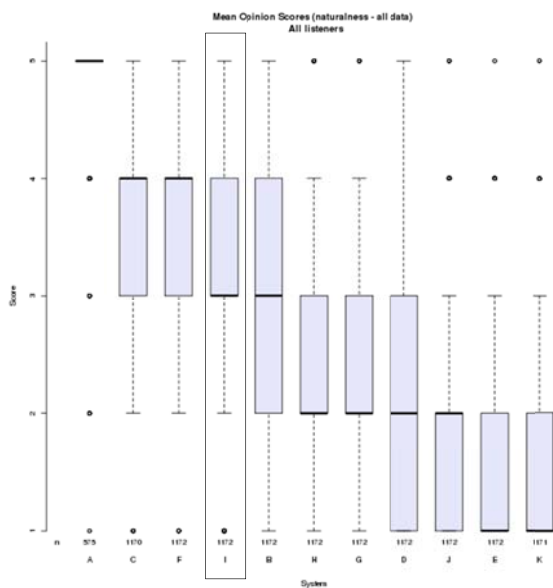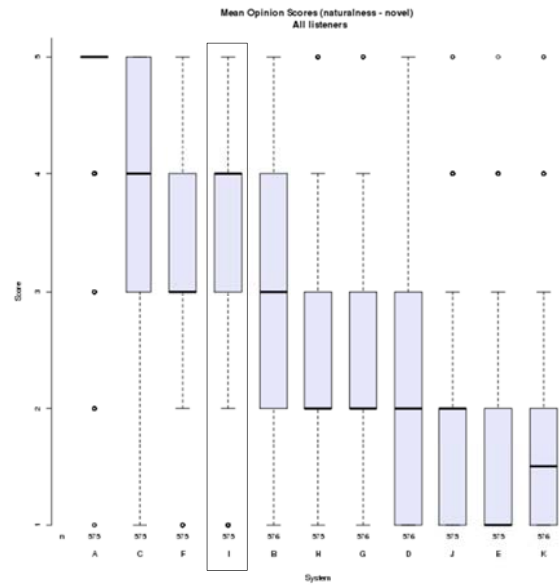Figure 2: *Mean opinion scores − naturalness (All listeners – All data).*

Figure 3: *Mean opinion scores − naturalness (All listeners – Novel stimuli only).*

4.1.2. Similarity to the original speaker

In the 'similarity to the original speaker' measure, our system got a mean score of 3.2 as depicted in Table 3. Although this result is better than last year's performance which is most probably attributed to specific improvements we have made in the concatenation and unit selection modules, we believe that the spectral equalization we performed on the original audio data may have affected negatively this task of the experiment.

Table 3. *Mean MOS-similarity-to-original-speaker scores for Blizzard Challenge 2012 for all participating systems.*

	All Listeners	Paid Listeners	Online Volunteers	Speech Experts
A	4,6	4,4	4,6	4,7
B	3,4	3,1	3,4	3,7
C	4,1	4	4	4,4
D	2	2	2	2
E	1,9	1,8	1,9	1,8
F	3,4	3,2	3,4	3,6
G	2,7	2,6	2,7	2,7
H	2,6	2,6	2,7	2,6
I	3,2	3,1	3,3	3,4
J	2,4	2	2,4	2,9
K	1,5	1,4	1,5	1,5

4.1.3. Word error rate (SUS experiment)

Regarding the word error rates (WER) this year's challenge included only SUS experiments, the scores of which are depicted in Table 4. Since this year we used both the

annotations and the phonetic transcription of the stimuli that were provided by the Blizzard organizing committee, there are no significant conclusions to make about language dependent modules such as the letter to sound one or the segmentation module. What is worth noting however is that this year's WER score is very similar to last year's score even though the training corpora are very different from each other [12].

Table 4. *Average Word Error Rate for SUS task in Blizzard Challenge 2012. All data and all listeners are depicted.*

	All Listeners	Paid Listeners	Online Volunteers	Speech Experts
B	26%	16%	38%	35%
C	19%	8%	33%	30%
D	23%	13%	33%	34%
E	49%	42%	58%	57%
F	27%	14%	39%	40%
G	26%	16%	37%	38%
H	23%	12%	34%	35%
I	24%	11%	38%	39%
J	39%	30%	47%	54%
K	32%	21%	42%	44%

Table 5. *MOS for different aspects of appropriateness of the participating TTS systems for audiobooks.*

	Overall Impression	Pleasentness	Speech pauses	Stress	Intonation	Emotion	Listening Effort
A	4,8	4,5	4,7	4,7	4,7	4,6	4,7
B	2,7	2,7	2,4	2,3	2,4	2,7	2,3
C	3,7	3,6	3,5	3,4	3,3	3,2	3,5
D	2,5	2,2	2,9	2,8	2,5	2,2	2,4
E	1,5	1,4	2,3	2,0	1,8	1,7	1,5
F	3,2	3,1	3,1	3,0	2,9	2,8	2,9
G	2,3	2,2	2,3	2,3	2,2	2,4	2,0
H	2,4	2,3	2,7	2,5	2,4	2,1	2,4
I	3,1	3,1	3,0	2,9	2,9	3,0	2,9
J	1,6	1,6	1,9	1,8	1,7	1,4	1,6
K	1,7	1,5	2,1	2,1	1,8	1,6	1,7

4.1.4. Appropriateness for Audio Books

The new section that was introduced in this year's Blizzard Challenge was the evaluation of the participating TTS systems when they are asked to cope with audio books and longer segments of text, such as entire paragraphs.

This experiment is quite innovative and it offers the opportunity to investigate the maturity of the participating

TTS systems against highly challenging tasks, as the audio book narration is.

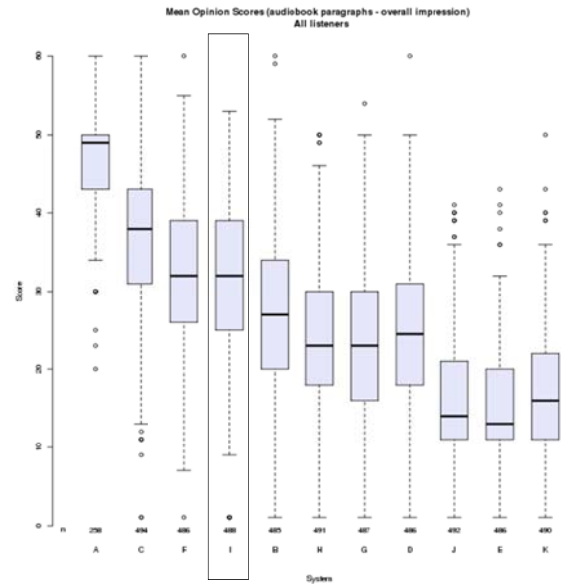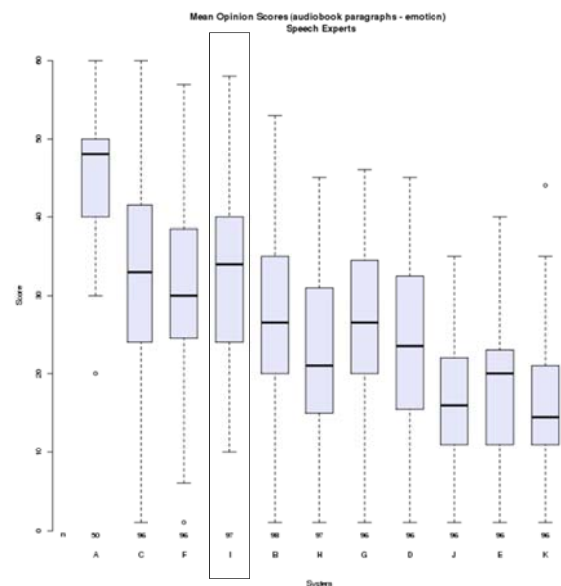Figure 3: *Mean opinion scores – Overall Impression All listeners.*

Figure 4: *Mean opinion scores − Emotion - Speech Experts.*

The aspects against which the TTS were evaluated are depicted in Table 5, where our system again is denoted with the letter 'I'. Our system ranked 3rd and 2nd in different aspects of the experiment, while in specific tasks that were completed by Speech Experts, our system ranked higher, like e.g. at the aspect Emotion where our TTS ranked 1st with a MOS 3.4 (Figure 3 and 4).

As far as the speech pauses and the stresses in the synthesized speech are concerned, we must clarify that we did not change either the pause lengths or stress annotations in the audio data provided. As already mentioned before, the prosodic annotation and intonation modeling during synthesis

we have used was our language independent algorithm which functions on phonetic level.

5. Discussion/Conclusions

One of our primary objectives for participating in this year's Blizzard Challenge was to put our voice building processes and tools to the test, and compare our progress in comparison to previous year's challenges. An additional reason however for this year's challenge was the idea of creating a TTS system with data that comes from an audio book, without any processing or supervision during recording. Creating a TTS system from raw data, and putting it into a demanding test, as reading an audio book is, was a challenge for us and a new field we would like to investigate further.

As a general outcome, our system's performance was improved in comparison to last year's participation (as far as similar experiment tasks are concerned). Improvements to concatenation and unit selection modules have been proven to affect positively our system's performance and efficiency. Core components of our system seem to be working equally well for different languages without significant adaptation (e.g. unit selection module, prosody generator) and with different speech domains, like for example expressive narration. The results depict that although there is a large room for improvement, the appropriateness of our TTS system for such use is considered to be acceptable. And by saying so, one can identify many different modules or algorithms of our TTS system that can be especially tweaked and improved for coping best with audio books, and those are many more than simply a richer prosodic modeling.

We believe that the area of expressive speech synthesis is still uncharted but this year's Blizzard challenge was one of the necessary steps speech synthesis needed to take towards it.

6. Acknowledgements

7. References

[1] Raptis, S. and Carayannis, G., "Fuzzy Logic for Rule-Based Formant Speech Synthesis," in Proc. EuroSpeech'97, Sept. 22-25, 1997, Rhodes, Greece

[2] Fotinea, S.-E., Tambouratzis, G., and Carayannis, G., "Constructing a Segment Database for Greek Time-Domain Speech Synthesis", in Proceedings of the Eurospeech-2001 Conference, Aalborg, Denmark, 3-7 September, Vol. 3, pp. 2075-2078.

[3] Karabetsos, S., Tsiakoulis, P., Chalamandaris, A., and Raptis, S., "HMM-based Speech Synthesis for the Greek Language" in Petr Sojka, Ivan Kopecek, and Karel Pala (eds.), 11th Int. Conf. Text Speech and Dialogue 2008 (TSD 2008), Book: Text, Speech and Dialogue, Book Series Chapter in Lecture Notes in Computer Science (LNCS), ISBN 978-3-540-87390-7, Springer – Verlag, Vol. 5246/2008, pp. 349 – 356

[4] Raptis, S., Tsiakoulis, P., Chalamandaris, A., and Karabetsos, S., "High Quality Unit-Selection Speech Synthesis for Bulgarian", In Proc. 13th International Conference on Speech and Computer (SPECOM'2009), St. Petersburg, Russia, June 21-25, 2009

[5] Karabetsos, S., Tsiakoulis, P., Chalamandaris, A., and Raptis, S., "Embedded Unit Selection Text-to-Speech Synthesis for Mobile Devices", IEEE Transactions on Consumer Electronics, Issue 2, Vol. 56, May, 2009

[6] Chalamandaris, A., Raptis, S., and Tsiakoulis, P., "Rule-based grapheme-to-phoneme method for the Greek", in Proc. Interspeech'2005: 9th European Conference on Speech Communication and Technology, September 4-8, Lisbon, Portugal, 2005

[7] Dutoit, T., "Corpus-based Speech Synthesis," Springer Handbook of Speech Processing, J. Benesty, M. M. Sondhi, Y. Huang (eds), Part D, Chapter 21, pp. 437-455, Springer, 2008.

[8] Karabetsos, S., Tsiakoulis, P., Chalamandaris, A., and Raptis, S., "One-Class Classification for Spectral Join Cost Calculation in Unit Selection Speech Synthesis", IEEE Signal Processing Letters, Vol. 17, No. 8, pp. 746-749, August, 2010

[9] Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P., "The HTK Book (for HTK version 3.2)", Cambridge University Engineering Department, 2002.

[10] Chalamandaris, A., Tsiakoulis, P., Karabetsos, S., and Raptis, S., "An efficient and robust pitch marking algorithm on the speech waveform for TD-PSOLA", 2009 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), vol., no., pp.397-401, 18-19 Nov. 2009

[11] Raptis S., Chalamandaris A., Tsiakoulis P.,Karabetsos S., "The ILSP Text-to-Speech System for the Blizzard Challenge 2010", In Proc. Blizzard Challenge 2010 Workshop, Kyoto, Japan, September 25, 2010

[12] Raptis S., Chalamandaris A., Tsiakoulis P.,Karabetsos S., "The ILSP Text-to-Speech System for the Blizzard Challenge 2011", In Proc. Blizzard Challenge 2011 Workshop, Torino, Italy, September 2, 2011