The ILSP Text-to-Speech System for the Blizzard Challenge 2010

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 the system employed by ILSP's Speech Synthesis Group for the Blizzard Challenge 2010 competition. It described the process of the building the required unit selection voices, and presents and discusses the obtained evaluation results.

Index Terms: speech synthesis, unit selection, speech evaluation

1. Introduction

This is the first participation of the Speech Synthesis Group of the Institute for Language and Speech Processing (ILSP), Athens, GREECE, to the Blizzard Challenge. This paper presents the system that the Group has used to enter the Blizzard Challenge 2010 competition.

ILSP has been in the forefront of 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 that Group employed for the Blizzard Challenge 2010 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 engine has been initially designed for the Greek language. However, as a corpus-based system, its design is, in most part, language-independent and has also been ported to the Bulgarian language with 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 prosodic and acoustic modules. In Section 3 the voice building process is explained. The evaluation results are presented and discussed in Section 4, and finally some conclusions are drawn in section 5.

2. System Overview

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

The two main modules that comprise 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 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.

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.

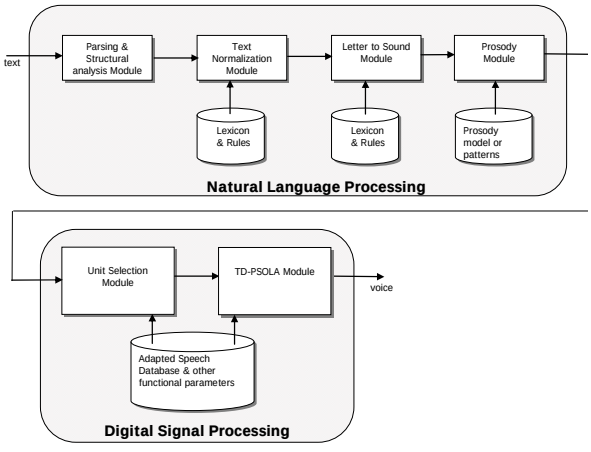

Figure 1: *Overall system architecture.*

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 has been 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 unit-selection. 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 the Time Domain Pitch Synchronous Overlap Add (TD-PSOLA) 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].

As typical in similar systems, the cost function employed in unit-selection is composed of two parts:

- *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 same values have been used for the weights of the cost function, which were manually-tuned for Greek. Due to time limitations, it was not possible to experiment with different values.

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

The typical Time Domain Pitch Synchronous Overlap Add (TD-PSOLA) method is used to concatenate the selected and enforce the smoothened pitch contour.

3. Building the Blizzard Voices

The following paragraphs describe the process of building the Blizzard 2010 voices for use with ILSP's TtS system.

3.1. Audio Preprocessing

Typically, the first of any voice building, is the amplitude normalization of the speech audio files. This was not necessary since the provided speech files were already normalized.

The version of the TtS engine used by the ILSP team worked internally at 44.1 KHz. We expected no appreciable improvement in the output speech quality to motivate the use of a higher internal sampling rate. Thus, the speech files which were provided at a 48KHz sampling rate were down-sampled before any further processing.

3.2. Building the Voices

This section provides a description of the steps we followed to build the Blizzard Challenge 2010 voices. Most parts of this procedure were similar for the English voices and for the Mandarin Chinese voice. Since the performance of our system with the Mandarin Chinese data was very poor, we do not go into details regarding its building procedure.

3.2.1. Labeling

For the phonetic and prosodic annotation of the speech corpus, we have not used the utterance files supplied. Instead, we chose to use our own custom label set which was the one also used in the letter-to-sound module.

Before arriving to that decision, some effort was first given in trying to align our custom phone set with the provided Festival phone set. That alignment stage was found to be much more demanding than originally estimated; the high inconsistency between the output of our letter-to-sound module and the provided labels made this mapping quite complicated and unreliable.

3.2.2. Segmentation

Since we did not use the labels provided, we had to perform segmentation of the entire corpus from scratch. To this end, we used the HTK [9] toolkit, followed by a set of custom post-processing scripts that identified and (to the degree possible) automatically corrected common segmentation errors.

Typically, an significant part of segmentation errors are related to breaths and inter-sentence pauses which are usually

not represented in the source text. However, since the majority of the sentences in the Blizzard corpus were rather small, such problems were quite rare.

Further to that, mismatches between the output letter-to-sound module and what was actually uttered, had an immediate, negative impact on the segmentation procedure.

Unfortunately, time constraints did not permit us to perform any thorough manual corrections.

3.2.3. Pruning

Due to time limitations, only automatic database pruning was performed. During this process database units that were considered to be outliers, based on specific pre-defined features such as duration, voiced/unvoiced switch etc., were excluded from the final database. By doing so, a maximum of 10% from the initial database was pruned.

3.2.4. Pitch-marking

For pitch marking, we employed the method described in [10].

4. Evaluation

As noted by the organizers, it is clear that all factors in all parts of the tests do NOT meet the normality requirements necessary to run parametric statistics. Additionally, most tests are carried out on an ordinal scale and the 'mean' and 'standard deviation' values are not meaningful. However, it is noted that a speculative ordering for the different systems can be extracted by ordering them by their mean MOS-naturalness score for the main voices. So such values are reported in the next paragraphs.

Furthermore, in order to provide some meaningful view and insight into the relative performance of our system, we include some mixed views that combine MOS-naturalness score and pairwise Wilcoxon signed rank test results, thus visualizing both relative ranking and significant differences between systems at the same time. These, along with the box-plots for MOS-naturalness score, provide a richer perspective of the competing systems.

The following sections summarize the results per voice. For each voice, results on similarity, naturalness and word error-rate are presented. Our system is identified by the "S" letter in the results files and plots distributed by the Blizzard organizers.

4.1. The rjs English Voice (EH1)

The rjs British English voice consisted of 5 hour (4000 utterance) database from a male professional speaker (RP accent) supplied by Phonetic Arts and available at 16kHz and 48kHz sampling rates, along with standard Festival labels produced by the University of Edinburgh. This voice was used for hub EH1 and for spokes ES2 and ES3. In the following, only the evaluation results for EH1 and ES3 are discussed.

Our system ranked at the 6th position in terms of the mean MOS-naturalness score among the 17 systems participating to EH1. It achieved a score of 3.1 with a standard deviation of 1.16.

Table 1 below, shows the Mean MOS-naturalness scores for EH1 for the benchmark systems, our system and the average score for all systems. For each system, mean scores are provided for all the listeners as well as breakdown information for paid (EE), volunteers (ER) and speech experts (ES) groups.

Table 1. *Mean MOS-naturalness scores for EH1 for the benchmark systems, ILSP's system and the average score for all systems. For each, mean scores are provided for all listeners as well as for paid (EE), volunteers (ER) and speech experts (ES) groups.*

System	All	EE	ER	ES
Natural speech	**4,80**	4,90	4,80	4,80
ILSP System	**3,10**	3,00	3,00	3,30
Festival Benchmark system	**3,00**	2,90	3,10	3,20
Average score of all systems	***2,86***	*2,78*	*2,88*	*3,03*
HTS_2005 Benchmark system	**2,50**	2,50	2,50	2,40

Figure 2 below shows standard boxplots for the Mean opinion scores for naturalness for task EH1 (all listeners).

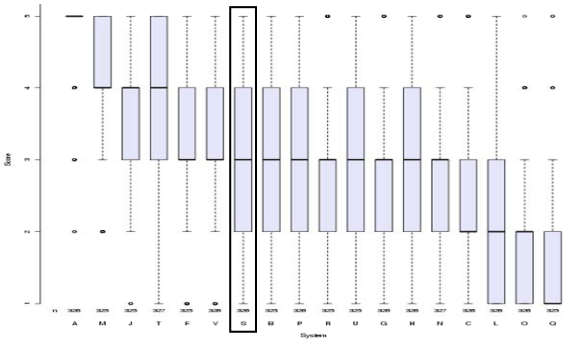

Figure 2: *Mean opinion scores − naturalness − for task EH1 (All listeners).*

Figure 3 shows a schematic overview of the performance of the various systems for EH1. Links between systems denote that there is no significant differences between them, according to pairwise Wilcoxon signed rank tests (at 1%

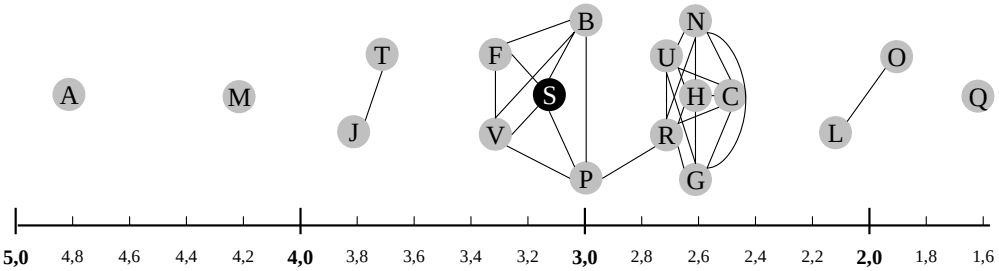

Figure 3: *Schematic diagram of the performance of the various systems for EH1. Links between systems denote that there is* no *significant differences between them, according to pairwise Wilcoxon signed rank tests(at 1% level).*

level). A look at this figure shows that our system does not have any significant difference from the two systems above it, namely, F and V. Likewise, it has no significant difference from the two systems following it, i.e. B (Festival Benchmark system) and P.

In the 'similarity to the original speaker' measure, our system got a mean score of 3.2. Interestingly, projecting this score to the different listener groups shows that speech experts provided significantly higher scores (3,6) than volunteers (3,1) and paid listeners (3,0).

Regarding the word error rates (WER) for the SUS test, our system performed quite poorly (0,26 mean score). However, only for the 6 out of the other 16 systems this difference is significant. Examining the WER results projected to different listener groups, one can observe large inconsistencies both in the score ranges and in the rakings that these imply for the different systems.

4.2. The 'Roger' English Voice (EH2)

The 'Roger' British English voice consisted of 1 hour (1000 utterance) subset of the 'Roger' database (male speaker, RP accent, ARCTIC sentences, same as in 2009) from the University of Edinburgh, with new hand-corrected labels supplied by iFLYTEK, along with standard Festival labels produced by the University of Edinburgh.

This voice was used for hub EH2 and, part of it (100 sentences) for spoke ES1. In the following, only the evaluation results for EH2 is discussed.

Our system ranked better than at the EH1, at the 4th position (along with system P) among the systems in terms of the mean MOS-naturalness score. It achieved a score of 3.1 with a standard deviation of 1.08.

Table 2. *Mean MOS-naturalness scores for EH2 for the benchmark systems, ILSP's system and the average score for all systems. For each, mean scores are provided for all listeners as well as for paid (EE), volunteers (ER) and speech experts (ES) groups.*

System	All	EE	ER	ES
Natural speech	**4,80**	4,90	4,80	4,80
ILSP System	**3,10**	3,10	3,20	3,10
Festival Benchmark system	**2,90**	2,80	2,90	3,00
Average score of all systems	***2,75***	*2,68*	*2,91*	*2,81*
HTS_2005 Benchmark system	**2,70**	2,80	2,70	2,50
HTS_2005 with hand-corrected labels	**2,60**	2,70	2,60	2,50

Table 2, shows the Mean MOS-naturalness scores for EH2 for the benchmark systems, our system and the average score for all systems. For each system, mean scores are provided for all the listeners as well as breakdown information for paid (EE), volunteers (ER) and speech experts (ES) groups.

Figure 2 below shows standard boxplots for the Mean opinion scores for naturalness for task EH2 (all listeners).

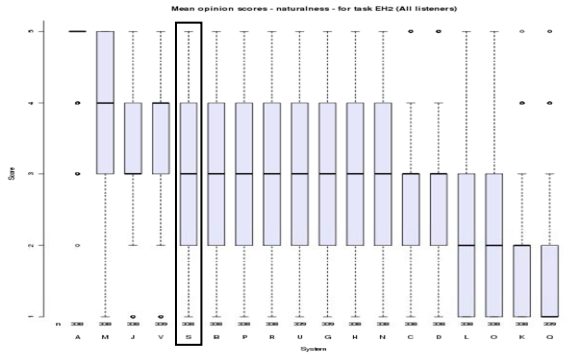

Figure 4: *Mean opinion scores − naturalness − for task EH1 (All listeners).*

Figure 5 shows a schematic overview of the performance of the various systems for EHS. As in Figure 3, links denote systems with no significant differences.

5. Discussion/Conclusions

One of our primary objectives was to put our voice building processes and tools to the test, by trying to avoid manual language-specific or voice-specific customizations as much as possible. This would provide some hints on how much of the prior information involved in voice building could be implicitly compensated for by relying on the corpus itself.

The system used by the group to enter the Blizzard Challenge 2010 for the British English voices was an adaptation of the group's TtS platform, employing customized tools by INNOETICS. Due to time restrictions, text analysis and prosodic modeling have been minimally addressed. Significant features such as POS-tagging, syntactic analysis and other characteristics that are also known to contribute to proper phrasing and are considered to be significant for English were not addressed in the system. As expected, this posed important limitations to the expected overall naturalness of the generated speech, especially in terms of prosody and speech flow.

Furthermore, custom labeling and segmentation of the speech files was performed almost fully automatically, without any manual corrections or manual database pruning.

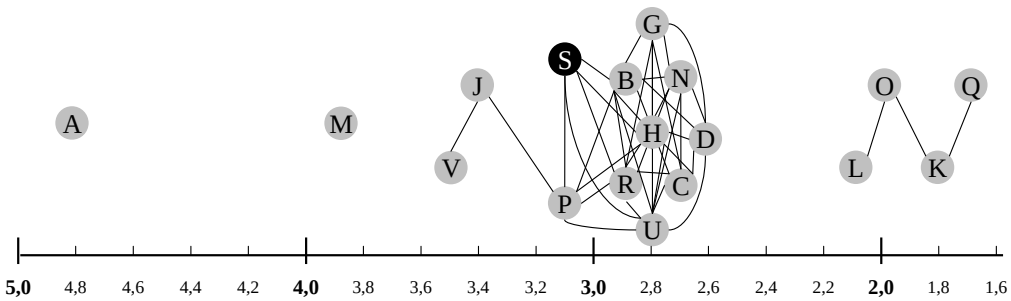

Figure 5: *Schematic diagram of the performance of the various systems for EH2. Links between systems denote that there is* no *significant differences between them, according to pairwise Wilcoxon signed rank tests(at 1% level).*

In this sense, this system is certainly not representative of the quality that can be achieved using our platform.

It is interesting to note that, as also reported in previous years, there seems to be an 'expert listener bias'. Breaking down the MOS-naturalness results by listener type reveals that the speech experts tend to provide higher scores than volunteers or paid listeners. This deviation was more evident for EH1 that it was for EH2. This is often attributed to that fact that people tend to get accustomed to listening to synthetic voices. One could also argue that the expectations of a speech expert from synthetic speech are different than those of a non-expert listener.

Regarding Mandarin Chinese, although the group did build a voice with the data provided, we do not present any information in the paper regarding the database creation and the evaluation results. This is due to the low performance achieved and to the fact that no customization of the modules and tools was actually performed. As a tonal language, Chinese clearly present significant differences from all the other languages our system has been tested upon. Several features that systems designed for other languages can afford to ignore, become much more relevant (and critical) when dealing with Chinese. Taking into account that this was the first time that the group addressed the Chinese language, the fact that no special customizations were made to the TtS engine, the time limitations, and the lack of a (native) Chinese speaker in the group, the poor performance of our system was not a surprise.

This first participation of the ILSP Speech Synthesis Group to the Blizzard Challenge has been a much enjoyed experience for us. We feel that such a competition is a great opportunity not only for understanding and comparing research techniques in building corpus-based speech synthesizers, but also for putting synthesis technologies, building procedures and speech tools to the test.

6. Acknowledgements

The authors would like to thank all the people involved in the organization and running of the Blizzard Challenge as well as the colleagues at ILSP who have offered their time for participating to the evaluation experiments.

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