The ILSP Text-to-Speech System for the Blizzard Challenge 2011

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 2011 competition. 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.

Index Terms: speech synthesis, unit selection, speech evaluation, Blizzard Challenge 2011

1. Introduction

This is the second 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 2011 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]), unit-selection, and statistical/parametric using HMMs [3].

The platform used for the entry is based on the core TtS engine by ILSP, as enhanced with speech tools and techniques by INNOETICS Ltd. It is very similar to the ILSP/INNOETICS entry for Blizzard 2010 which is described in [4]. The engine has been initially designed for the Greek language but has also been ported Bulgarian with high-quality results [5]. A scaled-down, low-footprint version of this system has also been developed for mobile environments [6].

This was our first US-English accented voice and therefore special customizations had to be performed during this year's challenge. The paper focuses only on the work required to support US-English, the necessary changes and adaptations, and the evaluation results.

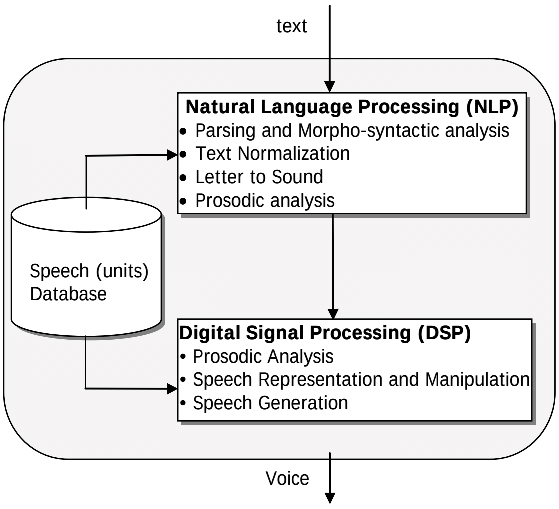
text
Natural Language Processing (NLP)
• Parsing and Morpho-syntactic analysis
• Text Normalization
• Letter to Sound
• Prosodic analysis
Speech (units) Database
Digital Signal Processing (DSP)
• Prosodic Analysis
• Speech Representation and Manipulation
• Speech Generation
Voice

Figure 1: *Overall system architecture.*

2. System Overview

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

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. This includes the letter-to-sound component where the technique employed was based on [7].

The DSP component comprises of the unit selection module which, as typical, is composed of a target cost component and a join cost component [8], and the signal manipulation module. The ILSP TtS system relies on a Time Domain Overlap Add method for speech manipulation. In adapting the synthesis engine to US-English, the weights for each component of the unit-selection cost function (many of which are phoneme-dependent) were manually tuned. 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.

3. Building a Voice from the Lessac Audio Data

The following paragraphs describe the process of building the Blizzard 2011 voices for use with ILSP's TtS system. The US-English voice for the Blizzard 2011 challenge was built using the provided ~15h long audio data. This data was provided to the Blizzard competition by Lessac Inc.

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.

3.2. Building the Voices

This section provides a description of the steps we followed to build the Blizzard Challenge 2011 voice. The same voice was used for all sections of the competition and no tailor-made voice was created for the task Addresses.

3.2.1. Labeling

For the phonetic and prosodic annotation of the speech corpus, we did not use the provided files or web services. Instead, we chose to use our own custom label set which was the one also used in the letter-to-sound module. As mentioned before, since this was our first attempt to build a US-English voice, we had to make specific customizations to the letter-to-sound rules in order to include US accent specific phenomena. Additionally to this, an exceptions lexicon was used in order to include words that were semi-automatically traced within the transcribed corpus and which were not properly addressed by the generic letter to sound rules.

3.2.2. Segmentation

In order to segment the audio data we used the HTK [9] toolkit, followed by a set of custom post-processing scripts that identified and automatically corrected common segmentation errors, identifying at the same time possible segmental errors.

The main source of segmentation errors were mismatches between the output letter-to-sound module and what was actually uttered. Also, as typical, a significant part of segmentation errors are related to breaths and inter-sentence pauses which are usually not represented in the source text.

No manual corrections or other supervised processing was performed during the segmentation process. Due to that, segmentation errors (wrong acoustic labels, wrong phoneme boundaries and/or misaligned pauses or breaths) were inherited by the database leading, in some cases, to poor performance.

3.2.3. Pruning

Due to time limitations, only automatic database pruning was performed. During this process, specific pre-defined features such as duration, voiced/unvoiced switch and spectral clustering were used as indicators of outliers, based on which sentences were excluded from the final database. By so doing, 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 Results

During Blizzard Challenge 2011, several aspects were put into evaluation. The main focus in our system is the level of naturalness achieved since we consider that as the dominant quality factor in a wide range of TtS applications. The similarity to the original speaker and the word error rate in SUS tests become important in specific application contexts.

Although most tests are carried out on an ordinal scale and the meaningfulness of the 'mean' and 'standard deviation' quantities may be rather limited, they were useful for us to gain an understanding of our system's performance and to obtain a relative ranking of our system compared to the other participating systems. Thus, a speculative ordering for the different systems can be extracted by ordering them by their mean MOS-naturalness score during the evaluation experiments.

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

4.1. The US English Voice (Lessac)

The Lessac US English voice consisted of approximately 16.5 hours (12,096 utterances) recordings from a female professional speaker supplied by Lessac Inc. and available at 16kHz and 48kHz sampling rates, along with Lessac labels produced by the aforementioned company. This same voice was used for all tasks, namely the evaluation tasks for similarity with the original speaker, for the level of naturalness and for the word error rate in SUS and Addresses subtasks.

Our system ranked at the 3rd position in terms of the mean MOS-naturalness score among the 12 systems participating to this task. It achieved a mean score of 3.2, while in the same test, for the listeners who were native English speakers our system ranked in the 2nd position with average MOS of 3.1.

Table 1 below, shows the Mean MOS-naturalness scores for this task, with an additional breakdown information for paid (EE), volunteers (ER) and speech experts (ES) groups, as well as native and non-native speakers. System A is natural speech. System B is a Festival benchmark system: this is a standard unit-selection voice built using the same method as used in the CSTR entry to Blizzard 2007. System C is a benchmark speaker-dependent HMM-based voice, built using a similar method to the HTS entry to Blizzard 2005.

Table 1. *Mean MOS-naturalness scores for Blizzard Challenge 2011 for all participating systems. For each, mean scores are provided for all listeners as well as for paid (EE), volunteers (ER) and speech experts (ES) groups.*

	All	EE	ER	ES	Native	Non-Native
A	4,7	4,6	4,6	4,8	4,6	4,7
B	2,7	2,4	2,7	2,9	2,5	2,8
C	2,7	2,6	2,9	2,8	2,6	2,9
D	2,6	2,3	3,0	2,7	2,4	2,8
E	3,3	3,0	3,4	3,6	3,0	3,6
F	2,5	2,2	2,9	2,6	2,3	2,8
G	3,9	3,6	4,0	4,1	3,6	4,1
H	3,2	2,9	3,3	3,4	3,1	3,4
I	1,5	1,5	1,5	1,4	1,5	1,4
J	2,4	2,3	2,3	2,5	2,4	2,4
K	3,1	2,8	3,2	3,3	2,9	3,3
L	3,0	2,6	3,2	3,3	2,7	3,3
M	2,7	2,4	3,1	2,9	2,5	2,9

In Figure 2 below, one can view the standard boxplots for the Mean opinion scores for naturalness for Blizzard 2011 (all listeners).

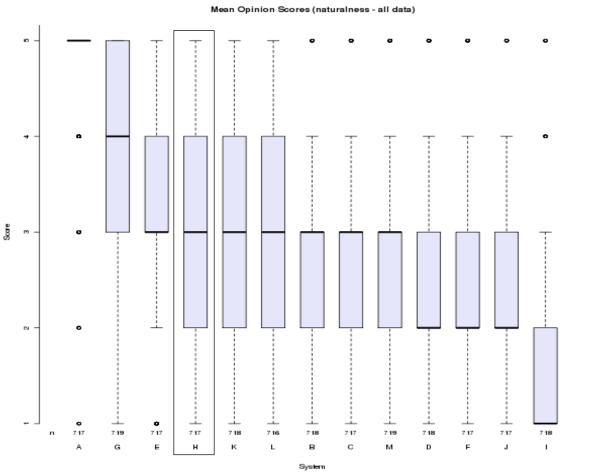

Figure 2: *Mean opinion scores − naturalness (All listeners).*

In the 'similarity to the original speaker' measure, our system got a mean score of 2.9. Table 2 below, shows the

Mean MOS-similarity-to-original-speaker scores for this task, with an additional breakdown information for paid (EE), volunteers (ER) and speech experts (ES) groups, as well as native and non-native speakers.

Regarding the word error rates (WER) for the SUS and Address tests, our system performed average (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.

5. Discussion/Conclusions

One of our primary objectives for participating for a second time to a Blizzard Challenge, was to put our voice building processes and tools to the test, correcting any problems or shortcomings identified during our previous participation [4]. Regarding out system, a general conclusion is that it ranked higher in comparison to last year's competition, thus validating our observations and consequent enhancements to the system and to the building process.

Table 2. *Mean MOS-similarity-to-original-speaker scores for Blizzard Challenge 2011 for all participating systems. For each, mean scores are provided for all listeners as well as for paid (EE), volunteers (ER) and speech experts (ES) groups.*

	All	EE	ER	ES	Native	Non-Native
A	4,8	4,8	4,6	4,8	4,8	4,7
B	2,9	2,7	2,8	3,0	2,8	2,9
C	2,6	2,6	2,4	2,6	2,6	2,5
D	2,4	2,3	2,6	2,4	2,4	2,4
E	3,1	3,0	2,9	3,2	3,0	3,2
F	2,4	2,3	2,4	2,4	2,4	2,3
G	3,3	3,0	3,1	3,6	3,2	3,5
H	2,9	2,8	2,7	3,1	2,8	3,0
I	1,4	1,4	1,6	1,3	1,4	1,4
J	2,5	2,5	2,4	2,7	2,5	2,6
K	2,8	2,7	2,8	3,0	2,8	2,9
L	2,8	2,7	2,7	3,0	2,7	2,9
M	2,7	2,6	2,8	2,8	2,7	2,8

Table 3. *Average Word Error Rate for both SUS and Address tasks for Blizzard Challenge 2011 for all participating systems. For each, mean scores are provided for all listeners as well as for paid (EE), volunteers (ER) and speech experts (ES) groups.*

	All	EE	ER	ES	Native	Non-Native
A	17%	3%	45%	23%	4%	32%
B	25%	11%	53%	33%	12%	42%
C	20%	7%	51%	26%	7%	36%
D	21%	7%	52%	26%	8%	36%
E	22%	9%	52%	28%	9%	38%
F	20%	7%	52%	26%	7%	37%
G	20%	7%	54%	25%	8%	36%
H	24%	11%	55%	31%	11%	41%
I	26%	12%	54%	33%	13%	43%
J	24%	10%	52%	31%	10%	41%
K	23%	9%	55%	30%	9%	40%
L	23%	11%	55%	28%	12%	38%
M	21%	7%	51%	27%	7%	38%

Improvements to our letter to sound module, as well as to the automatic segmentation process, seem to be significant in terms of the overall performance. Core components of our system appear to be working equally well for different languages without significant adaptation (e.g. unit selection module, prosody generator). Nevertheless, the main cause of mispronunciations or unnaturalness still lies in the letter-to-sound component and the segmentation process. Especially the first one is responsible for errors produced not only during the synthesis process but also during the segmentation process, since the phonetic labels on the recordings are produced by the same module.

The segmentation process on the other hand, is completely automatic and this allows errors to be propagated to the database building. Normally, a manual correction would give the "extra mile" in the performance of the system, but at a high effort cost. The database pruning which was performed offline, although it relieves the system from possible mismatches, does not solve the problem of bad segmentation or inconsistencies between the text script and the audio recordings.

Finally, significant features such as POS-tagging, syntactic analysis and other characteristics, which are also known to

Table 4. *Average Word Error Rate for SUS task for Blizzard Challenge 2011 for all participating systems. For each, mean scores are provided for all listeners as well as for paid (EE), volunteers (ER) and speech experts (ES) groups.*

	All	EE	ER	ES	Native	Non-Native
A	15%	5%	38%	19%	6%	27%
B	22%	10%	47%	27%	11%	35%
C	17%	8%	42%	21%	8%	29%
D	18%	8%	42%	22%	8%	30%
E	19%	9%	45%	24%	9%	32%
F	18%	7%	44%	22%	7%	31%
G	18%	7%	44%	22%	8%	30%
H	21%	10%	48%	25%	10%	34%
I	21%	11%	44%	27%	11%	35%
J	20%	9%	45%	26%	10%	34%
K	20%	9%	47%	25%	9%	34%
L	20%	10%	46%	24%	11%	32%
M	18%	8%	42%	23%	8%	31%

Table 5. *Average Word Error Rate for Address task for Blizzard Challenge 2011 for all participating systems. For each, mean scores are provided for all listeners as well as for paid (EE), volunteers (ER) and speech experts (ES) groups.*

	All	EE	ER	ES	Native	Non-Native
A	13%	9%	27%	13%	9%	18%
B	16%	9%	37%	18%	9%	24%
C	13%	9%	27%	12%	9%	17%
D	14%	9%	29%	15%	8%	21%
E	15%	8%	34%	17%	8%	23%
F	13%	7%	29%	14%	7%	20%
G	14%	8%	29%	17%	8%	21%
H	15%	8%	36%	16%	8%	22%
I	15%	8%	26%	19%	9%	22%
J	15%	8%	34%	18%	8%	24%
K	15%	9%	33%	17%	9%	23%
L	15%	8%	32%	19%	9%	23%
M	14%	9%	28%	15%	9%	21%

contribute to proper phrasing, were not addressed by our system. It is in our immediate future plans to address this issue, especially for English voices.

The algorithms and voice building processes used in ILSP and INNOETICS are constantly being improved and our participation to the Blizzard Challenge has been a much enjoyed and useful experience. 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, voice 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 and INNOETICS 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., 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

[5] 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

[6] 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

[7] 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

[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