The ILSP / INNOETICS Text-to-Speech System for the Blizzard Challenge 2014

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

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

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

Abstract

This paper describes the Speech Synthesis System entry for the Blizzard Challenge 2014 for INNOETICS and ILSP, along with the corresponding results and their analysis. We provide a description of the underlying system and techniques used in our TTS platform, as well as significant information about the voice building process. Based on the obtained results from the listening experiments, we attempt an evaluation of our system and the underlying methods. As this year's challenge included only Indian languages, we attempt to make a comparison with last year's results where the Blizzard Challenge introduced for the first time an Indian languages experimental hub. In the final section we provide a brief overview of our system's performance for the last three Blizzard Challenges, aiming mainly to identify its overall evolution.

Index Terms: speech synthesis, unit selection, speech evaluation, Blizzard Challenge 2013, innoetics, ILSP.

1. Introduction

For the Speech Synthesis Group of the Institute for Language and Speech Processing (ILSP), Athens, GREECE, and INNOETICS this was the fifth consecutive participation to the Blizzard Challenge. Although this year's challenge included only Indian languages, we happily participated and put our TTS platform into test, in order to investigate amongst others, how efficiently our system could handle new languages.

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]), unit-selection and an HMM parametric synthesis [3].

The system entry for the Blizzard Challenge 2014 is based on the core TtS engine developed by ILSP (unit-selection concatenative synthesis) with major enhancements and necessary speech processing tools, such as the voice creation platform developed by INNOETICS, a spin-off company offering commercial TtS solutions. Based on data-driven techniques, our system is a corpus-based TTS system and most of its modules are language-independent. As we have already successfully adapted it to support other languages such as Bulgarian and English [4], an important reason for participating to the Blizzard Challenges is to investigate how efficiently we can support new and often unknown languages within a short timeframe.

As the core components of our TTS platform remain the same with few additions or tweaks from time to time, we provide only a brief system overview here. A more detailed description of our TtS platform can be found in [5].

This paper is organized as follows. First, we describe the architecture of our system and 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 analyze them respectively.

2. System Overview

Following a front-end/back-end architecture, our TtS platform includes a Natural Language Processing (NLP) and a Digital Signal Processing (DSP) component, as illustrated in Figure 1.

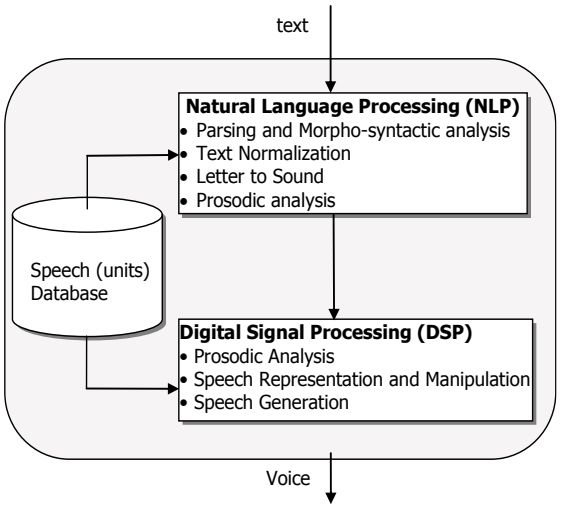

Figure 1: *Overall system architecture.*

2.1. The NLP Component (front-end)

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.

A detailed description of the underlying algorithms for letter-to-sound conversion and for modeling and reproducing the prosodic features of the recorded voice is provided in [6]. Since this year's challenge included Indian languages, an alternative approach has been taken for the letter-to-sound conversion, as this a language-dependent module.

2.2. The DSP Component (back-end)

The DSP component comprises of the unit selection module and the signal processing module, which 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.

The unit selection module is considered to be one of the most important components in a corpus-based unit selection concatenative speech synthesis system and 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] along with statistical calculation of the cost function features [8].
For concatenating and smoothening pitch discontinuities we have developed a custom pitch synchronous Time Domain Overlap Add (TD-OLA) method [9].

3. The Blizzard Challenge 2014

This year's challenge included two main tasks for 6 different Indian languages: Assamese, Indian, Rastathani, Tamil, Telugu and Gujarati. The basic challenge was to take the released speech data [10], build synthetic voices, and synthesize a prescribed set of test sentences. The output from each synthesizer was then evaluated through extensive listening tests. The first task included monolingual stimuli (native for each language) and the second task included the synthesis of multilingual stimuli (mixed with English text), based on the same voice data sets.
As the second task required knowledge of the languages, we participated only in the first task.
The training data provided was about 2 hours of speech data (sampled at 16 kHz) for each of the six aforementioned Indian languages, recorded by professional speakers in high quality studio environments. The text was provided in UTF-8 format with no other information, such as segment or phonetic labels.
For the IH1 task, a set of 150 stimuli was prescribed for each language, which later on was put into assessment by online listeners, both paid and volunteers, and strictly native speakers. During this task, both naturalness and similarity to the original speaker were evaluated, as well as word error rate in SUS experiments.

3.1. Building the IH1 Voices

For this task we employed the ILSP/INNOETICS TTS system using the automated voice building tool chain. The IH.1 task involved building voices for 6 Indian Languages with limited resources (about 2h of recorded speech and the corresponding script). The challenge for our team in this year's task was to put into test our automated voice building tool chain without having any native listener giving any feedback at any stage for any language, which was also the case during our participation in the Blizzard Challenge 2013.
The INNOETICS supporting tools that consist the voice building tool chain have been described in detail in the submissions for our entries in the previous Blizzard Challenges 2010 – 2013 [7,11,12]. These tools highly automate the voice building procedure for any language, given a text processing front-end. Since, we had no such front-end for any of the specific Indian Languages; our main work involved the investigation of basic approaches.

3.1.1. Data Preparation

A set of recorded sentences was provided for each of the IH languages corresponding to each language. Although the voice data was recorded in good studio conditions, we performed a normalization along with a filtering and equalization stage of the audio files in order to alleviate intensity and spectral mismatches. Since all recordings were of similar quality we did not exclude any of the data provided for training our TtS platform.

3.1.2. Pitch-marking

For pitch marking, we utilized the method we have developed and which is described in [17].

3.1.3. Front-End

The letter-to-sound component is a core requirement in the front-end module and since we had no knowledge of the target languages, we investigated two basic approaches:

a) by using a letter based approach, i.e. the alphabet of each language becomes the phone set, and each letter becomes a unique phone.

b) by using a third party tool for text processing. For the latter we used the eSpeak synthesizer [13] for letter to sound conversion, at least for the supported languages.

For each language we built two different synthetic voices, a letter-based and a phone-based one. In order to choose between each pair of synthetic voices for each language we performed a short informal listening test using a very small set of sentences held out from the training data. The sentences were synthesized with each voice and compared against the original wave file with the help of the Amazon Mechanical Turk Service. The results showed that the phone based systems for Hindi, Tamil and Telugu were slightly better than the corresponding letter-based ones and hence we decided to use the phoneme-based voices for Hindi, Tamil and Telugu and the letter-based voices for Assamese, Rajasthani and Gujarati.

3.1.4. Segmentation

For segmenting the audio data we used the INNOETICS voice production tool-chain which is based on an HMM forced-alignment algorithm [14,15]. The alignment has been performed without any change or supervision as it closely developed to the TTS front-end component, using the same resources and modules for the text-processing stage.

3.1.5. Pruning

We performed an automatic pruning of the segmented audio recordings mainly based on two features: a) the relative duration for each label and b) the relative HMM score derived during the forced-alignment between the text and the audio. In this year's task, the training material was better arranged since it was designed and produced for use within a TtS platform, and therefore we did not performed any pruning stage for removing possible mismatches between text and audio, as this was often the case in previous year's challenge where the training material was audiobooks.

3.1.6. Back-End

The back-end processing modules in our system are in general language independent and required no further adaptation for the IH tasks.

4. Evaluation Results

4.1. The IH.1 Tasks (monolingual)

Similarly to previous Blizzard Challenges, in this year's challenge three main aspects were put into evaluation via listening tests: naturalness, similarity to the original speaker and word error rate in SUS sentences. The listening subjects

were native speakers for each language as each listener had to go through a language dependent CAPTCHA test in order to complete a task for a specific language.

In the following results our system is identified with the letter "G", while "A" is the natural speech.

4.1.1. Naturalness

As far as the naturalness of the synthetic stimuli is concerned, our system performed exceedingly well and received the best MOS score in all six languages. Especially for the Assamese and Telugu the difference between our system's performance and the second's best was statistically significant.

Table 1. The overall results for IH tasks (Naturalness – all listeners – Mean Opinion Score). In Assamese and Telugu the difference with the second best is statistically significant.

	Assamese	*Gujarati*	*Hindi*	*Rajasthani*	*Tamil*	*Telugu*
A	4,7	4,7	4,5	4,2	4,6	4,9
B	2,1	2,6	2,0	2,3	2,3	2,0
C	3,3	2,8	2,5	3,3	2,7	3,1
D	3,5	2,8	3,6	3,7	3,2	3,5
E	2,9	3,5	3,1	3,7	2,9	3,1
F	3,4	3,4	3,2	3,9	3,4	4,0
G	3,9	3,8	3,7	3,9	3,6	4,2
H	-	2,5	2,2	3,1	2,7	1,9
I	2,1	2,7	2,2	3,2	2,6	2,3
J	-	-	-	-	2,6	-
K	-	-	3,4	-	-	-

As naturalness is maybe the most important aspect of a TTS system, after intelligibility, our system seems to capture it efficiently, with a relative high MOS score, especially when taking into account that the distance from the natural utterances is generally less than one on a five-unit scale.

4.1.2. Similarity to the original speaker

As far as the similarity to original speaker is concerned, our system performed also very well and received the best MOS score in four out of the six languages. Especially for the Tamil and Telugu the difference between our system's performance and the second best was statistically significant.
In our case, any slight differences from the original speaker may have be attributed mainly to the DSP filtering of the wave files and secondly to the limitations of the time-domain manipulation methods.

Table 2. The overall results for IH tasks (Similarity to the original speaker – all listeners – Mean Opinion Score). In Tamil and Telugu the difference with the second best is statistically significant.

	Assamese	*Gujarati*	*Hindi*	*Rajasthani*	*Tamil*	*Telugu*
A	3,3	2,9	4,3	4,4	4,0	4,5
B	1,8	3,0	2,4	2,6	2,0	1,7
C	2,8	3,0	2,6	3,5	2,6	2,6
D	3,2	2,7	4,0	3,6	3,0	2,5
E	2,6	3,5	3,2	3,6	2,7	2,3
F	2,9	2,8	3,4	4,0	2,7	3,3
G	3,2	3,7	3,4	3,7	3,8	3,9
H	-	3,5	2,1	3,1	3,2	1,4
I	1,8	2,8	3,1	3,3	1,8	2,9
J	-	-	-	-	3,1	-
K	-	-	2,4	-	-	-

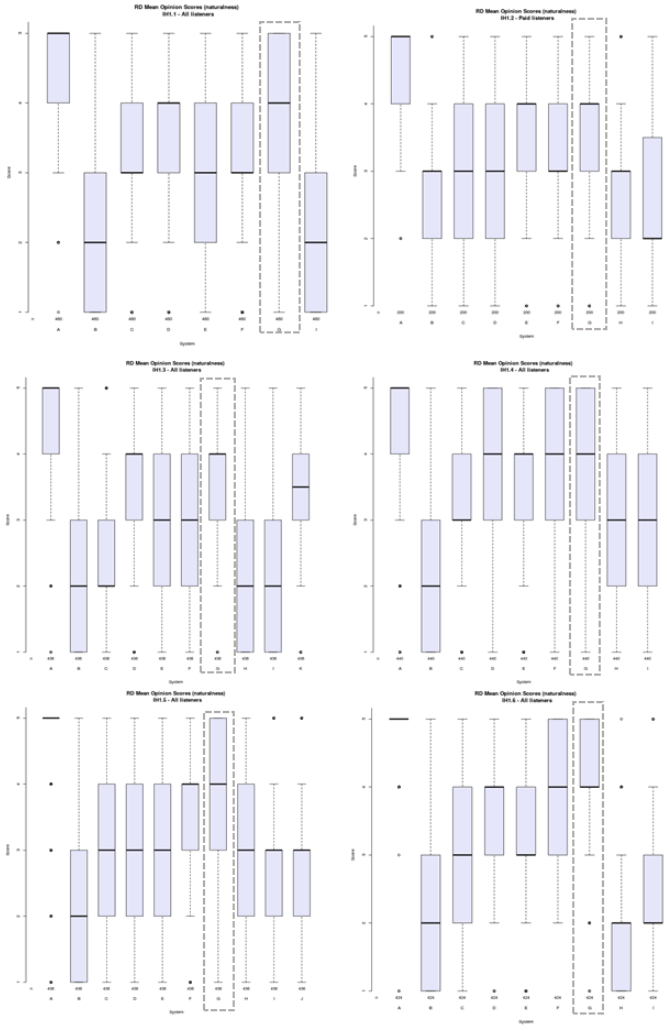

Figure 1: The Median values and their standard deviation of the MOS scores for each language for the evaluating naturalness (all listeners). The dashed-lined box depicts our system's performance.

4.1.3. Word Error Rate

As far as the Word Error Rate of the SUS synthetic stimuli is concerned, our system performed better than the average in most languages. At this point however, and by observing

rather high WER in most languages, even for the natural stimuli, one must note that this specific task is rather difficult to perform as there are inconsistencies and not standardized methods for text input for the target languages. This issue had been identified during Blizzard Challenge 2013 with an attempt to partially solve it by linking the experiment page with the Google Transliteration service [16].

Table 3. The overall results for IH tasks (Word Error Rate – all listeners – SUS sentences).

	Assamese	*Gujarati*	*Hindi*	*Rajasthani*	*Tamil*	*Telugu*
A (Natural)	0,51	0,24	0,22	0,62	0,32	0,40
B	0,86	0,34	0,26	1,00	0,33	0,55
C	0,84	0,59	0,40	0,67	0,64	0,77
D	0,69	0,40	0,24	0,65	0,38	0,54
E	0,76	0,23	0,27	0,60	0,37	0,51
F	0,67	0,25	0,24	0,64	0,37	0,46
G	0,74	0,37	0,29	0,59	0,37	0,51
H	-	0,41	0,30	0,67	0,60	0,57
I	0,69	0,44	0,30	0,57	0,34	0,62
J	-	-		-	0,44	-
K	-	-	0,25	-	-	-
Mean Val.	0,75	0,38	0,28	0,67	0,43	0,57

5. Discussion/Conclusions

Like in previous years, our primary objective for participating in the Blizzard Challenge this year was to put our voice building processes and tools to the test, and compare our progress in comparison to previous year's challenges. As this year's challenge included languages on which there was no knowledge in the team, we wanted to investigate how efficiently our voice building and TTS platform can perform in totally new languages. The creation of new unknown languages is a challenge for us and for our future plans.

As a general outcome, our system's performance was exceedingly high as it received the top score in all six languages as far as the naturalness is concerned. This is the second year in a row that our system receives the best MOS score in Indian languages, similarly to the Blizzard Challenge 2013. In comparison to the results of our participation in 2013 (as far as Hindi and Tamil are concerned) our system seems to have performed better this year in Hindi, since the baseline is the same (same MOS of the natural utterance). For Tamil we cannot reach to a firm conclusion as our MOS score this year is slightly lower, but with a higher MOS score of the natural utterance, meaning that the higher quality audio recording lead to higher expectations from the synthetic stimuli.

This year, our system received an average MOS score of 3.9 for all six languages which is significantly better than last year's performance. Aside from the improvements we have integrated into our voice building and TTS platform, this improvement in the score can be partially attributed to the higher quality of the original recordings in 2014, compared to 2013. These recordings, although limited in length, were well designed and produced for use with a TTS engine.

Regarding the similarity to the original speaker, our system performed very well receiving the top score in four out of the six languages, and the main reason for deviating from the top MOS score in Hindi and Rajasthani was most probably the DSP filtering stage we carried out during the audio preparation (normalization and equalization).

As far as the WER task is concerned, our system had a better than the average performance; nevertheless the fact that high WER values were observed for the natural utterances for most of the target languages, hints that the text input method in the experiments may have affected the overall scores.

Building new voices and new languages for a TTS platform is one of the most important processes, which require however a lot of effort and resources. The Blizzard Challenge is a great tool for evaluating and improving this tool chain and process.

6. Acknowledgements

The authors would like to thank all the people involved in the organization and running of the Blizzard Challenge. The research leading to these results has been partially funded by a Greek National GSRT funded project (GSRT Project Code 54NEW_B_2012).

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] Tsiakoulis, Pirros, et al. "An Overview of the ILSP Unit Selection Text-to-Speech Synthesis System." Artificial Intelligence: Methods and Applications. Springer International Publishing, 2014. 370-383.

[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] Raptis, Spyros, et al. "The ILSP Text-to-Speech System for the Blizzard Challenge 2012." Proc. Blizzard Challenge 2012 Workshop, Kyoto, Portland, Oregon USA. 2012.

[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] 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.

[10] Patil, Hemant A., et al. "A syllable-based framework for unit selection synthesis in 13 Indian languages." Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2013 International Conference. IEEE, 2013.

[11] 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

[12] 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

[13] Duddington, Jonathan. "eSpeak Text to Speech." (2010).

[14] Braunschweiler, Norbert, Mark JF Gales, and Sabine Buchholz. "Lightly supervised recognition for automatic alignment of large coherent speech recordings." Proceedings of the 11th Annual Conference of the International Speech Communication Association. Curran Associates, Inc., 2010.

[15] 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.

[16] Prahallad, Kishore, et al. "The Blizzard Challenge 2013–Indian language task." Blizzard Challenge Workshop 2013. 2013.