The ILSP / INNOETICS Text-to-Speech System for the Blizzard Challenge 2013

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

[1] Institute for Language and Speech Processing / Research Center "Athena", Athens, Greece
[2] INNOETICS LTD, Athens, Greece
{aimilios,ptsiak,sotoskar,spy}@ilsp.gr

Abstract

This paper describes ILSP and INNOETICS Speech Synthesis System entry for the Blizzard Challenge 2013. A description of the underlying system and techniques used is provided, as well as information about the voice building process and discussion on the obtained evaluation results. Extra focus will be given on the new techniques we used this year in comparison to our previous participations, and we will also attempt a comparative analysis of this year's results with the results of the Blizzard Challenge 2012, aiming to investigate the abilities and the progress performed as far as expressive speech synthesis is concerned.

Index Terms: speech synthesis, unit selection, speech evaluation, Blizzard Challenge 2013, audio books, librivox, expressive speech synthesis.

1. Introduction

This is the fourth consecutive participation of the Speech Synthesis Group of the Institute for Language and Speech Processing (ILSP), and INNOETICS to the Blizzard Challenge. This paper presents the system used for the ILSP/INNOETICS entry to the Blizzard Challenge 2013 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]), unit-selection and an HMM parametric synthesis [3].

The system entry for the Blizzard Challenge 2013 is based on the core TtS engine by ILSP and enhanced with speech tools and techniques by INNOETICS, a spin-off company offering commercial solutions based on the core technology. It is a corpus-driven TTS system and most of its modules are language-independent, with already successful migrations and customizations to other languages such as Bulgarian and English, offering equally high-quality results [4]. A scaled-down, low-footprint version of this system has also been developed for mobile environments [5]. The core technological modules underneath our TTS platform remain the same with few additions or tweaks from time to time; therefore we shall provide a brief description of the underlying technology for the sake of completeness, as it has been already been published in previous Blizzard Challenges reports.

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 analyze them respectively.

2. System Overview

Like most TTS systems, our platform follows a typical concatenative, unit-selection architecture as depicted in Figure 1, with a Natural Language Processing (NLP) and a Digital Signal Processing (DSP) component collaborating in the heart of the system.

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. This year, we incorporated an additional functionality of POS tagging (Part-of-Speech) which was further used in the unit-selection algorithm. A brief description of every sub-module is given below:

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-

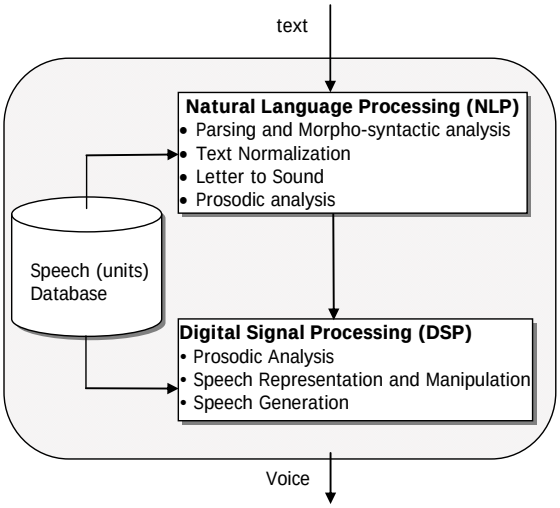

Figure 1: *Overall system architecture.*

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

A new feature in the 2013 entry for the EH task is the addition of part-of-speech information in the prosody model used in the 2012 entry [7]. The prosody is modeled implicitly in a data driven manner. 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. Prosody is modeled in terms of target pitch values or duration models taking into account the distance of the unit 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.
This model was further enhanced to include part-of-speech information in the distance scoring metrics. More specifically the prosodically salient unit set was extended to include the part-of-speech tag of the containing word. The cardinality of the resulting set is the original cardinality multiplied with the number of distinct part-of-speech tags used. The abundance of the available data (especially for the EH1 task) makes the coverage of salient unit combinations in the database more probable. A small scale listening test showed a preference towards the new model compared to our previous approach.

2.2. The Acoustic Subsystem

The DSP component comprises of the unit selection module and the signal manipulation 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. A brief description of these modules is 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 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].
For the target cost two 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. As mentioned above, this year we added a POS cost component to this cost as brief tests illustrated an improvement in the performance of the unit-selection module.
For the join cost two 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.

3. The Blizzard Challenge 2013

This year's challenge included two main tasks, one for the English language and one for Indian languages (EH and IH tasks). For the EH task, audiobook data was kindly provided by The Voice Factory, from a single female speaker, provided as approximately 300 hours of chapter-sized mp3 files, plus approximately 19 hours of non-compressed wav files. The wave files had been segmented into sentences and aligned with the text by Lessac Technologies, Inc. In the EH1 subtask, the participants were asked to build a TTS system with the entire audio data provided, and in the EH2 subtask a TTS system with the 19 hours audio data only. The text books for the audio data could be found in Gutenberg project as the original books were free of rights.
For the IH task, four different subtasks were planned. A set of 1000 sentences was provided for 4 different Indian languages, namely Hindi, Bengali, Kannada, and Tamil, and the participants were asked to create a TTS system for each data set.
For each subtask of the Blizzard Challenge, synthetic stimuli were put into assessment by online listeners, both paid and volunteers, and by speech experts. Different aspects of the synthetic speech were asked to be rated in every subtask.

3.1. Building the EH1 and EH2 Voices

The following paragraphs describe the process of building the Blizzard 2013 EH voices for our TTS system. In both EH1 and EH2 subtasks, we used the original audio data and performed the necessary segmentation with our own tools; hence for the EH2 task we did not use the segmented wave files provided by Lessac, but the original audiobook data for each book. All audiobooks were narrated by the same voice talent, however some of the audiobooks provided were recorded in different environmental settings and therefore we had to process them or even exclude them from the database.

3.2. Audio Preprocessing

The first step was the selection of the appropriate audiobooks to include in the database, or more precisely selecting the audiobooks to exclude from the rest of the material, as they would probably cause spectral discontinuities due to different environmental recording settings. To do so, we performed a clustering of the audio files, based on the SFFT spectral content of the recordings. Based on random samples within each audio book, our algorithm clustered the Audiobooks data

into 3 different clusters, which we validated via manual listening of the samples. We decided to exclude 8 out of the 30 audiobooks provided as their recording settings were different from the rest of the audio data. All audio data was converted into 44KHz wave files and then processed for segmentation and labeling.

3.2.1. Labeling

For the phonetic annotation of the speech corpus, we used our letter to sound and prosodic modules for the US-English phonetic annotation and prosodic features respectively. An additional POS labeling took place but on the phonetic level of the recordings, after the phonetics segmentation of the audio data.

3.2.2. Segmentation

The main issue that needs to be addressed for the creation of a database from an audiobook is the alignment of the audio recordings with the actual spoken text and any other annotation is necessary on the audio segments [8].
As we use an HMM alignment mechanism, as described in [9] between the audio and the text parts of the audiobooks, we performed an iterative process of aligning them, using smaller and smaller parts of the recordings each time, in order to increase the accuracy of the alignment. Starting from large chunks of speech, ranging from 20 minutes to 1,5h long (most of the times chapters or bigger sections of an audiobook), we ran the audio alignment again on a second phase on a sentence level, after the initial alignment. This approach provides better and more accurate results [10]. The alignment has been performed without any significant change or supervision of the input text, meaning that possible mispronunciations or disagreements between the text and the recorded speech could exist. Most of these parts will be excluded automatically at a later stage of the database crafting process, during pruning.

In the table below, one can see the audiobook data used in the EH tasks after the segmentation and pruning stages.

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

	EH1	*Length*	*Further*	*DB Length*
1	black	4:42	35%	3:10
2	mansfield	13:08	32%	8:51
3	awakening	4:47	33%	3:15
4	black	4:36	32%	3:10
5	chatterley	10:18	31%	7:12
6	cityoz	0:42	30%	0:29
7	daisy	1:01	29%	0:40
8	dalloway	4:22	33%	2:58
9	emerald	0:42	32%	0:29
10	emma	14:05	30%	9:26
11	frankenstein	6:32	33%	4:22
12	jane	16:13	33%	11:40
13	leagues	8:24	28%	5:57
14	madding	12:32	29%	8:46
15	magi	0:12	30%	0:08
16	patchwork	2:09	31%	1:27
17	pride	11:08	32%	7:27
18	roomview	6:50	33%	4:38
19	scarlet	6:18	32%	4:17
20	treasure	6:13	32%	4:28
21	washington	6:00	28%	4:07
	Total	140:54:00	31%	97:07:16

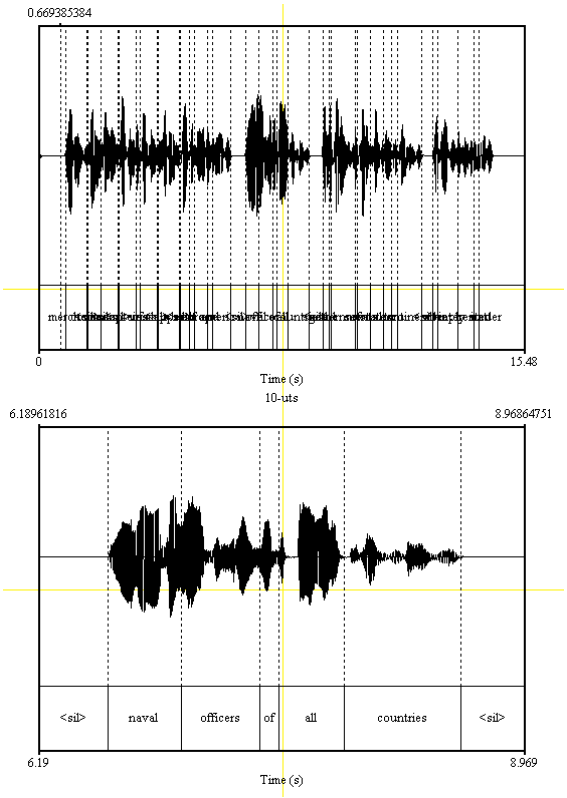

Figure 1: The segmentation and alignment of the audio data is performed in an iterative method, starting from long audio chunks (e.g. chapters or sections) and concluding to a phrase level, as it identified between two consecutive recognized speech pauses. Phrases are then aligned with the text and annotated on a phonetic level.

3.2.3. Pruning

We performed an automatic pruning of the segmented audio recordings aiming on two goals: a) the pruning of segmental errors and b) the pruning of different speech styles within the recordings.

3.2.3.1 Segmental errors pruning

During the automatic alignment process between the audio recordings and the actual text of the audiobooks there are segmental errors which are caused either by possible disagreements between the text and the spoken version of it, or by possible fails of the actual aligning mechanism. In order to address this issue, a simple but efficient stage of pruning was introduced, which depending on the local HMM score of every phoneme and the overall score of every sentence, would make a decision on whether an aligned sentence would be appropriately segmented and annotated, and hence a good addition to the TTS database. By doing so, sentences or words that had received low score by the HMM algorithm during alignment were removed from the recordings pool and the voice database crafting process. The decision threshold for rejecting a sentence or a word was estimated manually and this process stage led to the exclusion of 31 percent of the available recordings.

3.2.3.2 Prominent speech pruning

As already mentioned, audiobooks are an exciting material for text-to-speech voice crafting as they include various linguist

and acoustic phenomena, which can lead to emotional text-to-speech systems if modeled and reproduced appropriately. However, since our aim was to build a more generic TTS voice, we attempted to identify and remove most prominent speech parts of the recordings from the database. The issue of identifying different speech styles within audio recordings has been addressed previously [11-14] with different results depending on the set of acoustic and linguistic features extracted from the audio and text respectively.

Our methodology employs two acoustic features only, the mean and the variance of the F0 variable, on each phrase. By phrase we mean the audio part between two consecutive recognized silences or pauses by the voice alignment engine, and not a textual sentence, as the narrator tends to keep the same speaking style within a phrase, and make a pause before changing speaking style. From the entire population of the recognized phrases, we keep only a portion (based on a threshold) of the phrases which is located closest to the centroid of the distribution based on a specific distance. The threshold is again estimated manually and depends not only on the narrator (e.g. if he/she likes to play different roles with different voices and so on), but also on the actual book, as a book may contain more dialogues or roles than others.

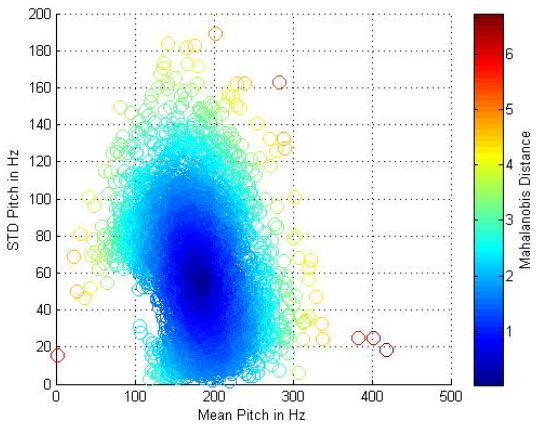

Figure 2: Distribution of the acoustic phrases based on their Mean Pitch value versus the Standard Deviation Pitch value. The Mahalanobis distance of each instance from the population's centroid is annotated via a color map.

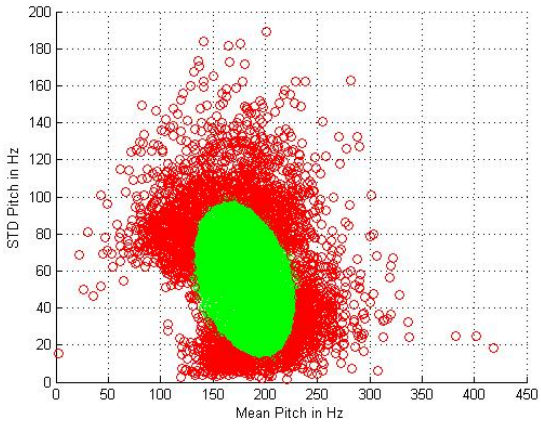

Figure 3: The phrases that were selected to be included in the database are denoted with a green circle. The selection criterion is based on a manually estimated threshold of the calculated distance.

3.2.4. Pitch-marking

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

3.3. 1. Building Voices for the IH Tasks

Next we describe the process of building voices for the IH task for the ILSP/Innoetics TTS system using the automated voice building tool chain. The IH tasks involved building voices for four Indian Languages with limited resources. The challenge for our team for this task was to put into test our automated voice building tool chain without having any native listener giving any feedback at any stage.

The ILSP/Innoetics TTS system and supporting tools have been described in detail in the submissions for our entries in the previous Blizzard Challenges 2010 – 2013 [7,18,19]. 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 Indian Languages; our main work involved investigating basic approaches.

3.3.1. Data Preparation

A set of 1000 recorded sentences were provided for each of the IH tasks corresponding the following languages: Hindi, Bengali, Kannada, and Tamil. The data were recorded in various conditions (often having background noise, high reverberation, far-field speech, etc.) which posed an additional challenge to the task. This was more evident for the Bengali data set. We tried noise removal and sound restoration techniques in the degraded speech samples, but the result was not satisfactory. The residual speech samples were still highly mismatched with the rest of the data, so we decided to discard them from the speech repository. We excluded the sentences ben_0451 to ben_1000 from the Bengali speech database, and the sentences kan_0422 to kan_0500 from the Kannada one.

3.3.2. Front-End

The letter-to-sound component is a core requirement in the fron-end module. For this, we investigated two basic approaches:

a) use a letter based approach. The default implantation of the letter-to-sound component is pass-through, i.e. the alphabet becomes the phone set, and each letter becomes a phone.

b) use a third party tool for text processing. For the latter we used the eSpeak synthesizer [20] to convert text into phonetic text followed by a simple mapping into our phone set.

We built two voices for each of the IH tasks using the letter based as well as the phone based approach without any other task specific refinement. To choose between each pair of voices an informal listening was carried out 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. This showed that the phone based systems were slightly more accurate. A brief investigation of the mismatches for the letter based case showed that most of the problems were due to diacritic-like symbols that do not correspond to phonemes, but rather alter the surrounding ones. For this we decided to enter the IH tasks using the phone based voices.

3.3.3. 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 EH Tasks

Like in Blizzard Challenge 2012, in this year's challenge several aspects were put into evaluation for the EH tasks. For the assessment of the TTS systems in coping with books, seven different aspects were tested in total: overall impression, pleasantness, speech pauses, stress, intonation, emotion and listening. Furthermore, the listeners were asked to assess the performance of the systems, as far as their intelligibility (Word Error Rate), the similarity to the original speaker and their performance in other types of text such as news, novel are concerned.

In the following results our system is identified with the letter "L", while "A" and "B" are the natural speech and the festival benchmark system accordingly.

4.1.1. The EH MOS Tasks

In the EH1 subtask, we developed a database from all available audio data, after having discarded 8 out of 30 audiobooks, while in the EH2 subtask we used the audio material from only 2 audiobooks. The overall MOS results for each subtask are shown in the following table and the overall impression is shown in the following figures.

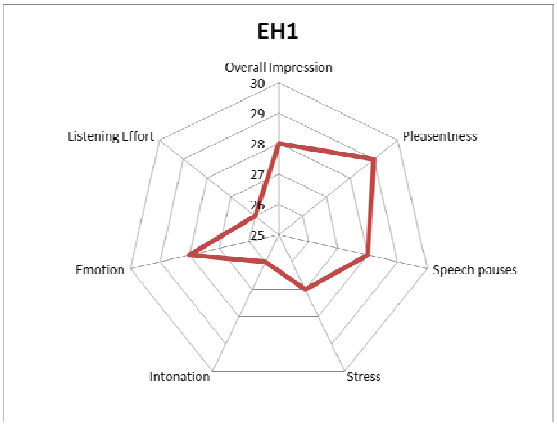

Figure 4: The MOS score in different aspects of speech assessment for our system. Task EH1.

Our system performed exceedingly well, especially in the EH2 subtask where it was rated first in the overall impression criterion among all listeners.

If we wanted to attempt a comparison to last year's results for our system, we could see an improvement in almost every field. One should note here that the data sets between this and last year's Challenges were rather different and it could be possible that a comparison would not be appropriate.

Table 2. The overall results for EH1 task (paragraph) for all systems and all listeners.

	Overall Impression	*Pleasantness*	*Speech pauses*	*Stress*	*Intonation*	*Emotion*	*Listening Effort*
A	4,7	4,6	4,6	4,7	4,6	4,6	4,7
B	1,7	1,7	1,9	1,9	1,8	2	1,5
C	2,7	2,5	3	2,9	2,6	2,4	2,7
F	1,7	1,7	2,2	2	1,9	1,7	1,8
H	1,8	1,7	2,4	2,1	1,7	1,3	1,9
I	2,7	2,5	3,1	3	2,7	2,4	2,7
K	3	3	3,2	3	3	2,8	3
L	2,8	2,9	2,8	2,7	2,6	2,8	2,6
M	3,6	3,5	3,4	3,3	3,2	3,3	3,3
N	2,2	2,3	2,1	2,1	2,1	2,1	2,1
P	1,1	1,1	1,5	1,2	1,1	1,1	1

Table 3. The overall results for EH1 task (paragraph) for all systems and all listeners.

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

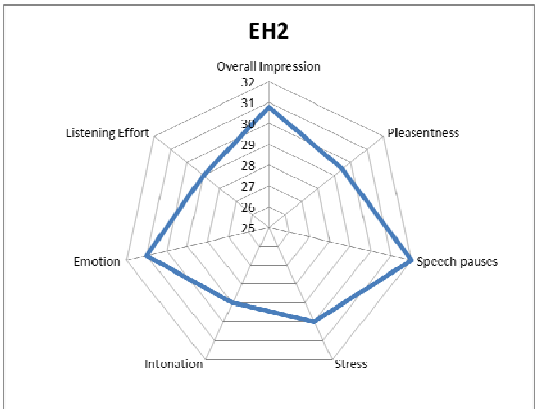

Figure 5: The MOS score in different aspects of speech assessment for our system. Task EH2.

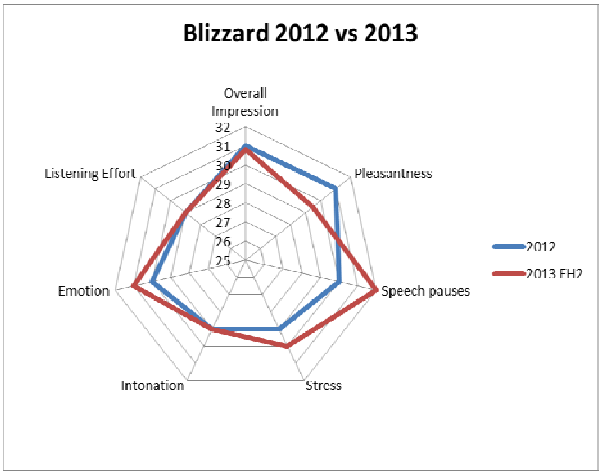

Figure 6: The MOS score in different aspects of speech assessment for our system. Task EH2 2013 with task EH 2012 for the ILSP/INNOETICS TTS system.

4.1.2. The EH SUS and SIM tasks

Our system performed also very well in the MOS tests where all TTS systems were assessed in news and novel sentences. The stimuli were assessed for both naturalness and similarity to the original speaker. Our system was rated among the first systems in both subtasks.

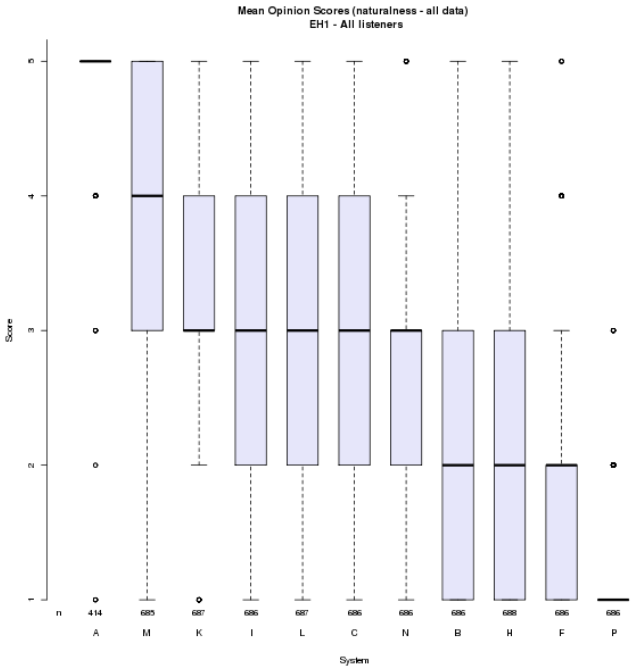

Figure 7: MOS on naturalness for the task EH1 for all listeners and all data. Our system is depicted with the letter L.

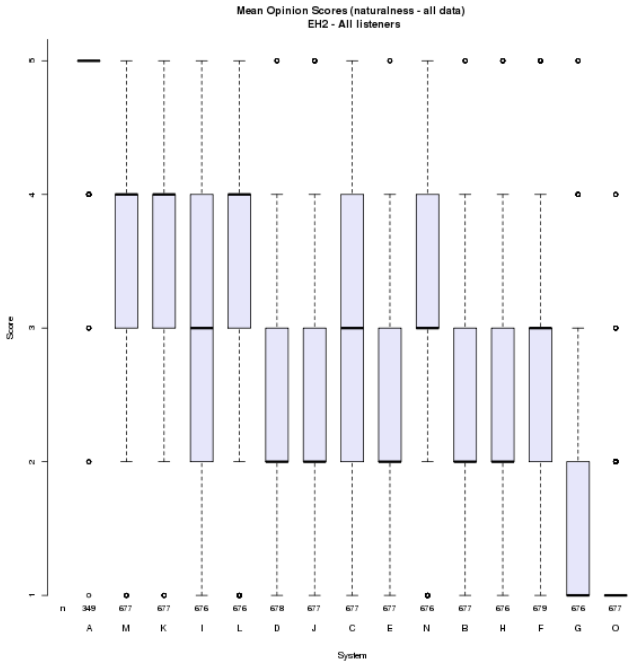

Figure 8: MOS on naturalness for the task EH2 for all listeners and all data. Our system is depicted with the letter L.

As far as the SUS subtasks are concerned, our system performed about average with a considerable high Word Error Rate; this can be attributed entirely to the insertion of segmentation errors during the database crafting.

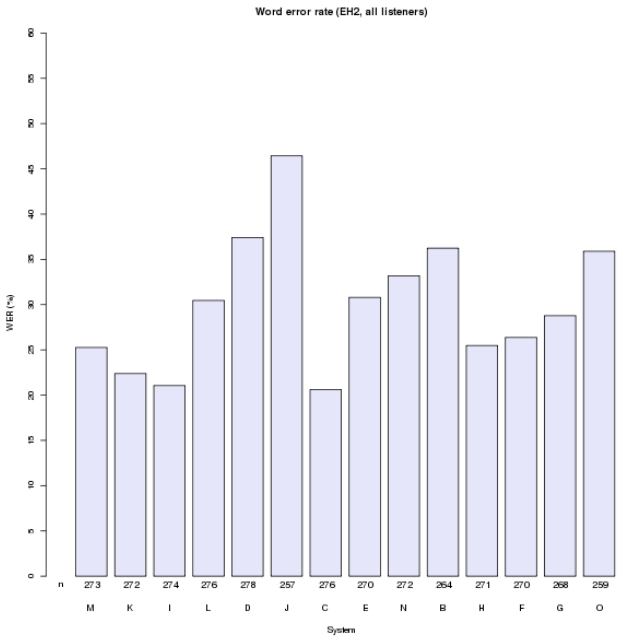

Figure 9: Word error rate for the task EH2 for all listeners and all data. Our system is depicted with the letter L.

4.2. The IH Tasks

This year's challenge included a pilot task for assessing the TTS systems in 4 Indian languages, namely Hindi, Bengali, Kannada, and Tamil. The assessment of the stimuli focuses on the naturalness and the similarity to the original speaker, as well as on the word error rate.

In both metrics for similarity to the original speaker and naturalness, our system was rated first in all subtasks, with significant difference from the rest, in most cases.

Table 4. The overall results for IH tasks for our system. All data and all listeners are included.

	Similarity	*Naturalness*
IH1	3,0	3,6
IH2	3,4	3,8
IH3	2,5	3,7
IH4	3,3	3,8

One should note here that although the training data was very limited the results were exceedingly good but for the subtask IH3 where the data set included very poor recordings.

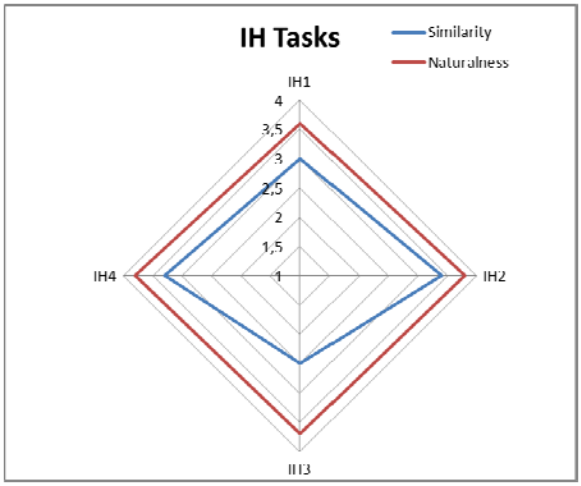

Figure 10: The MOS score in different aspects of speech assessment for our system. Tasks IH1-4 for Hindi, Bengali, Kannada, and Tamil respectively.

As far as the intelligibility metric of the participating systems was concerned, the resulting average WER depicts that the training data set was insufficient as the average WER for all systems among all listeners was over 50%.

5. Discussion/Conclusions

As in 2012, our primary objective 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. The creation of synthetic voices from audio books is a very challenging but also promising task, as the crafting of a synthetic voice is most of the times an expensive in time and money part of a TTS system.

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) with our system been ranked in the top TTS systems in all aspects of the assessment, while in many aspects it was ranked first. Improvements to concatenation and unit selection modules have been proven to affect positively our system's performance and efficiency as well as the addition of the POS labeling of the acoustic units. By participating in the pilot tasks for Indian languages we concluded that the core components of our system seem to be working equally well for different languages without significant adaptation (e.g. unit selection module, prosody generator) as our system ended-up in the first position in both similarity and naturalness.

As the training material is not designed for a TTS voice crafting, many issues have to be resolved during this process, or otherwise the derived voice database will provide low-quality results, with inconsistent and often non-intelligible speech. These issues include: a) segmentation errors introduced by several factors such as text versus voice disagreements, mispronunciations or voice mimicking and role playing by the narrator; b) the existence of different speaking styles within the recordings and c) the possible different environmental recording settings that would cause the recordings to have different audio quality. Although our methodology for addressing these issues seems to work efficiently enough, there is still room for improvement in all the aforementioned stages (e.g. segmentation, pruning, equalization). We plan to work more on the pruning techniques in order to improve both intelligibility and overall performance of our TTS system. We believe that this year's Blizzard challenge was again a good step towards better speech synthesis.

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 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] 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] 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.

[10] 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.

[11] 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.

[12] Wang, Lijuan, et al. "Exploring expressive speech space in an audio-book." Proc. of Speech Prosody 2006.

[13] Székely, Eva, et al. "Detecting a targeted voice style in an audiobook using voice quality features." Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. IEEE, 2012.

[14] Charfuelan, Marcela, and Marc Schröder. "Correlation analysis of sentiment analysis scores and acoustic features in audiobook narratives." ES3-LREC Workshop. 2012.

[15] Chen, Langzhou, et al. "Exploring Rich Expressive Information from Audiobook Data Using Cluster Adaptive Training."

[16] Székely, Eva, et al. "Clustering expressive speech styles in audiobooks using glottal source parameters." Proc. of Interspeech, Florence (2011): 2409-2412.

[17] 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

[18] 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

[19] 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

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