Expressive Speech Synthesis for Storytelling: The INNOETICS' Entry to the Blizzard Challenge 2016

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

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

Abstract

This paper describes INNOETICS' Speech Synthesis System entry for the Blizzard Challenge 2016, along with the corresponding results and some relevant discussion. We provide a description of the underlying system and techniques used in our TTS platform, as well as some detailed information regarding the voice building process. Based on the obtained results from the listening experiments, we attempt an evaluation of our system and the underlying methods.

Index Terms: expressive speech synthesis, hybrid TTS system, speech evaluation, Blizzard Challenge 2016, innoetics

1. Introduction

This was the sixth participation of INNOETICS to the Blizzard Challenge and one of the most challenging ones as it involved involving audiobook content for children storytelling. INNOETICS is a spin-off company from the Institute for Language and Speech Processing / "Athena" Research and Innovation Center, which has been at the forefront of text-to-speech R&D in Greece for almost two decades, having developed TtS engines for the Greek language based on all the major approaches: from formant rule-based systems (e.g. [1]), to diphone (e.g. [2]), unit-selection (e.g. [3]) and to HMM parametric synthesis [4].

INNOETICS' current TTS technology is based on a state-of-the-art hybrid TTS engine that combines the merits of data-driven modeling with the near-natural quality achieved through waveform concatenation. Coupled with a strong speech processing toolset, it offers an efficient voice building pipeline that can deliver rich, top-quality synthetic voices in short timeframes.

The back-end of the TTS engine and the voice building pipeline are language-independent and have been successfully applied to a range of languages, delivering commercial-grade voices for languages such as English, Greek, Bulgarian [5], Arabic and Russian, while several additional languages are currently in beta. In addition, it has been used to develop synthetic voices for 6 Indian languages, all of which have ranked at the top of the Blizzard Challenge 2014 [6].

One of the main challenges of the Blizzard Challenge 2016, was the need to cope with rich, expressive speech from audiobooks addressed to young children. This placed considerable burden both to the TTS engines and to the voice building processes as they needed to effectively handle a range of issues such as highly colored speech, voice character imitations, non-linguistic vocalizations and audio effects which quite frequently appeared in the content.

Although identifying and discarding such segments can be a viable strategy in cases where a 'neutral' synthetic voice is desired [7], the same was not true in the case of the Blizzard Challenge, as the resulting voice would be used to synthesize content from the same domain: children stories. Thus, preserving the richness of expressive styles and developing strategies for invoking them as needed was an indispensable part of the task.

Furthermore, the large breadth of phenomena that were present in the content leaded to a high dimensional expressive space where even the 5-hours speech data of the provided audiobooks seemed sparse, making it difficult for unsupervised clustering approaches (e.g. [8, 9]) to deal with.

This paper is organized as follows. First, we describe the architecture of our system in section 2 and in section 3 we describe the voice building process and specific adaptations that were necessary for this challenge. The evaluation results obtained by our system are presented in section 4. Section 5 includes some relevant discussion. Finally, section 6 includes some more general comments regarding the research task of designing expressive / emotional speech synthesis systems.

2. System Overview

The overall architecture of our TTS engine follows a typical front-end/back-end layout, comprising a text-processing component and a signal processing component, as illustrated in Figure 1.

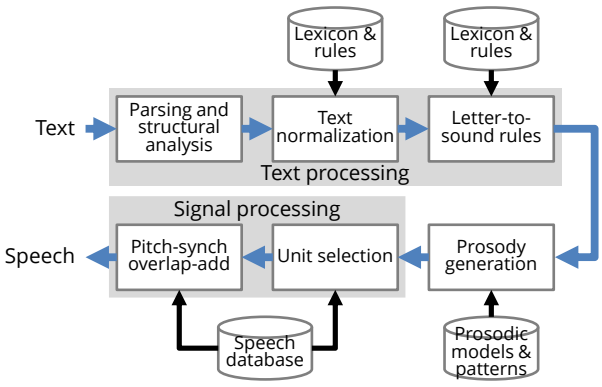

Figure 1: *Overall architecture of the TTS engine*

Diphones are the main units of our system, but there is a mechanism for falling back to demi-phones when a diphone is

missing or a "good-enough" instance of the diphone is not available in the database for a specific context.

2.1. The front-end

The text-processing component is mainly responsible for parsing, analyzing and transforming the input text into an intermediate format appropriate to feed the DSP component. This involves extracting the necessary features from the text which specify the phonetic, linguistic and prosodic targets for each unit and drive the unit selection procedure that follows.

This year's challenge involved English which are already supported by our system. So, we employed the sentence- and word-tokenization, text normalizer and letter-to-sound modules that we had already available.

A few additions were made to the **pronunciation dictionary** in order to ensure that proper names that appeared in the training and test data were correctly phonetized. A few short audio extracts were provided by the Blizzard Challenge organizers to illustrate the pronunciations of some words in the test data set. These were only used to clarify the pronunciation and update the lexicon and not included in the training data.

The front-end is also responsible for providing essential information regarding **prosody**. Our system does not employ any explicit prosodic model in the form of specific target values for pitch, duration or intensity. Instead, it takes into account the prosodic context of a unit as part of its target cost (see next section about the back-end). So, it is part of the front-end's tasks to extract such prosodic information.

An important customization that was introduced to the front-end of our system for this year's challenge was the concept of two "**domains**": the "`narration`" domain which comprised all utterances that were not part of a dialogue and the "`character`" domain which included all dialogues. Practically, the front-end relied solely on the existence of quotation marks on the input text to decide which of the two domains to assign to each text segment. Mixed sentences were treated accordingly by splitting them into segments and assigning different domains to each, as in the following example:

"Listen to this," he announced to his friends.
| `character` | `narration` |

The underlying hypothesis was that the dialogue part of the stories would tend to be significantly more expressive than the narrative parts, such as in the case of voice acting for imitating the voices of specific characters in a story. In general, this distinction proved to be quite effective in controlling the overall quality of speech and ensuring a smoother rendering. However, a closer analysis of the acoustic features of the two groups revealed a considerable overlap in the acoustic space between the two categories. Practically, this means that it was not too uncommon for a narration part to be quite expressive (i.e. the narration often got quite loaded), or, on the opposite side, for a dialogue part to be less so (e.g. a dialogue or a specific character in a story were spoken in a rather neutral style).

2.2. The back-end

The DSP component comprises of the unit selection module which performs the selection of appropriate units from the speech database, and the signal processing module which relies on the time-domain pitch synchronous overlap and add method for speech manipulation and concatenation.

Unit-selection is one of the most important modules in a concatenative speech synthesis system and its performance directly affects the quality of the synthetic speech. Our system falls under the "hybrid" unit-selection category, as some of the components of the cost function rely on statistical information and models that are calculated during the voice building stage. These components relate both to the target cost as well as to the join cost [10].

So, the unit selection module involves the minimization of a total cost function which comprises of two partial cost functions, namely the target cost and the concatenation cost function.

There are three components in the **target cost**. One that accounts for the similarity of the phonetic context, one that accounts for the graphemic context (for instance, whether a unit is at a word boundary or not), and one that accounts for the prosodic context (i.e. a unit's distance from significant neighboring prosodic boundaries, such as stressed syllables and punctuation). Part of the target cost function was also the "domain", as discussed in the front-end (Section 2.1). Based on the front-end processing, each unit in the database was assigned one of the two domains: either `narration` or `character`. When synthesizing `narration` parts, unit-selection only used `narration` units. However, for synthesizing `character` utterances, unit selection favored `character` units but did not completely exclude `narration` units. This strategy was dictated both by intuition and for practical reasons as the `narration` segments in the database were nearly twice as many as the `character` segments.

The **join cost** mainly takes into account the spectral and pitch continuity of the two units, as well as the duration of the two adjacent demiphones of the diphones to be merged.

3. The Voice Building Process

This year's challenge involved about 5 hours of speech data from professionally-produced children's audiobooks along with the corresponding text scripts. All recordings were from a single speaker. The task was to build a synthetic voice from this data that would be suitable for reading audiobooks to children.

3.1. Data preparation and cleanup

Some of the books were provided in PDF format. We converted them automatically to text through OCR. The conversion introduced some errors which were manually corrected through spell-checking and inspection.

In addition, a script was used to balance quotation marks in the text and the result was verified/corrected through inspection. This step was important as quotation marks were used by our system as the main means to split the training corpus into the `narration` and `character` domains.

Finally, the labels of some non-linguistic vocalizations (e.g. "argh") were edited to a "phonetically similar" form. Although no effort was devoted into making these forms consistent or appropriate for any further processing, they did provide a rough representation of the vocalization that helped the segmentation process work its way around these vocalizations.

The speech recordings of the audiobooks were not provided in a single audio format but included a mix of MP3, M4A and WMA files. These were all converted to WAV format to be used

for segmentation. All the recordings were of similar quality so there was no need for excluding any part of the content.

Some of the audiobooks used a non-vocal "ding" sound as a page separator, i.e. to signal the listener to turn page in the physical book. An acoustic model trained on this sound was used to remove them so that they did not interfere with the segmentation process.

3.2. Sentence-level alignment

The original sentence-level alignment provided with the original data was not used. Instead, we have re-aligned the speech data at the sentence level based on the corrected text scripts (balanced quotation marks and labels of non-linguistic vocalizations) and preserving as much of the punctuation as possible. The resulting lab files were uploaded to the corresponding git repository that had been setup by the organizers in order to share them with the other participants.

3.3. Segmentation, labeling and pruning

For segmenting the audio data at the phone-level we used the INNOETICS' voice production pipeline which is based on an HMM forced-alignment algorithm [11, 12]. The alignment was performed without supervision and it employed the same front-end component as the one used for synthesis thus ensuring a consistent behavior between the building and synthesis stages.

The "phonetically similar" forms that we used as labels for some of the non-linguistic vocalizations helped the segmentation process work its way around these vocalizations. So, at the end of the segmentation process we had information for the boundaries of these vocalizations. Part of these were kept so that they were available during synthesis, but not as part of the regular unit selection process. Rather, they were made available as "emoticons" that the user could embed "as-is" in an utterance at synthesis time through appropriate notation.

In overall, only a very limited part of the data provided in the training dataset were pruned and excluded from the speech database. These related to non-vocal sounds, non-linguistic vocalizations, overacted speech / extreme character imitations. These were mainly identified by the low scores that the respective units received during segmentation and were verified through inspection.

The segmentation stage also assigned phonetic and prosodic labels at the phone level using custom label sets, as well as domain labels (`narration` vs. `character`).

3.4. Audio analysis and feature extraction

An analysis stage followed the segmentation process, where a set of acoustic features where calculated for the identified units. This included the features necessary for calculating concatenation join costs at synthesis runtime such as, for instance, spectral and prosodic measures at the unit boundaries.

For pitch marking, we utilized the method we have developed and which is described in [13].

4. Evaluation and Results

The details of the evaluation procedure is described in the Blizzard Challenge call while more details will be given in the summary Blizzard Challenge paper by the organizers.

The evaluation comprised of a number of sections, some of them involving entire paragraphs of synthetic speech and some single sentences. Including the original voice (natural speech), there were 17 systems involved in the evaluation (16 for SUS as there was no natural speech for these).

The performance of our system in each evaluation section is discussed in the paragraphs below.

4.1. Audiobook paragraphs

Listeners listened to a whole paragraph from a children's book and chose a score on a scale of 1 to 60 for different aspects of the synthetic speech. Fig. 2 provides an overview of the performance of the different systems for comparison.

Based on a speculative ordering of the participating systems based on their mean score, our system ranked at the second position. However, as shown in Table 1 below, a statistical analysis of the responses of the speech experts and the paid listeners shoed no statistically significant difference between the INNOETICS system and the top-ranked system for any of the evaluation dimensions. Only for the online volunteers and for some of the evaluation dimensions was the difference significant.

Figure 2: *Overview of the MOS of the participating systems in different dimensions for all the listeners. Values are in the range 10-50. The green line is the natural voice and the red line is the INNOETICS system. The rest of the systems are shown in light gray, while the average of all the participating systems is shown in black.*

Table 1. *Statistical significance between the INNOETICS system and the top-ranked system*

	Paid listeners	Online volunteers	Speech experts
overall impression	NO	YES	NO
pleasantness	NO	YES	NO
speech pauses	NO	YES	NO
stress	NO	NO	NO
intonation	NO	NO	NO
emotion	NO	NO	NO
listening effort	NO	YES	NO

4.2. Audiobook sentences - Naturalness

In this listening test, the subjects listened to one sample and graded it on a scale of 1 ("Completely Unnatural") to 5 ("Completely Natural"). Fig. 3 summarizes the performance of all participating systems while Table 2 provides detailed values.

A speculative ordering of the participating systems ranked our system at the second position while it has no statistical significance from the highest-ranked system as determined by Wilcoxon's signed rank tests using a Bonferroni correction.

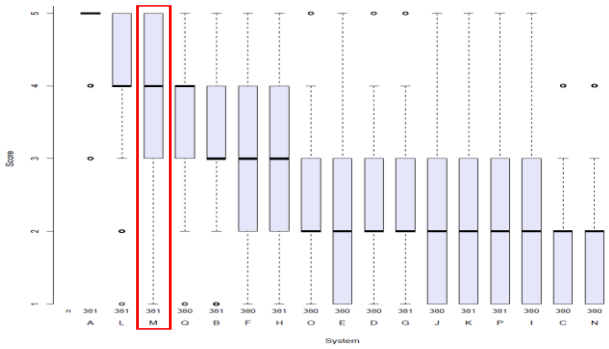

Figure 3: *Overview of the MOS of the participating systems for naturalness among all the listeners. Values are in the range 1-5. System A is the natural voice. The performance of the INNOETICS system (letter M) is marked in red.*

Table 2. *Detailed information regarding the naturalness of the INNOETICS system in audiobook sentences.*

	Median	MAD	Mean	Std Dev
Natural voice	5	0	4,8	0,42
INNOETICS	4	1,5	3,9	0,9
Average			2,72	

4.3. Similarity to original speaker

In this listening test, the subjects could play 4 reference samples of the original speaker and one synthetic sample. They chose a response that represented how similar the synthetic voice sounded to the voice in the reference samples on a scale from 1 ("Sounds like a totally different person") to 5 ("Sounds like exactly the same person").

Fig. 4 summarizes the performance of all participating systems with regard to the similarity to the original speaker while Table 3 provides detailed values for our system.

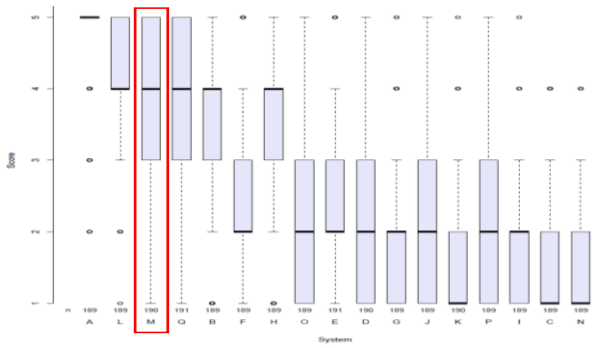

Figure 4: *Overview of the MOS of the participating systems for the similarity to the original speaker among all the listeners. Values are in the range 1-5. System A is the natural voice. The performance of the INNOETICS system (letter M) is marked in red.*

Table 3. *Detailed information regarding the naturalness of the INNOETICS system in audiobook sentences.*

	Median	MAD	Mean	Std Dev
Natural voice	5	0	4,7	0,65
INNOETICS	4	1,5	3,9	0,93
Average			2,64	

Ordering of the participating systems based on their mean score ranks our system at the third position regarding the similarity to the original speaker, but with no statistical significance from any of the two higher-ranked systems.

4.4. Semantically unpredictable sentences

Listeners heard one utterance in each part and typed in what they heard. Similar sentence types were used as in the previous year. Listeners were allowed to listen to each sentence only once. The word error rate was computed as in previous years.

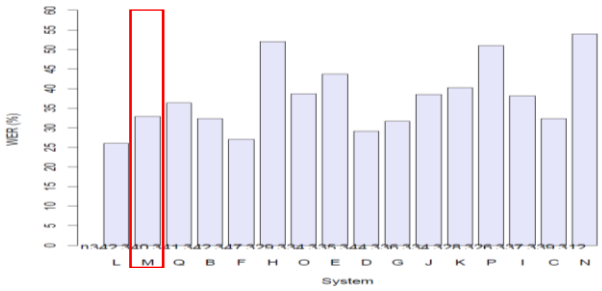

Figure 5: *Overview of the word error rate (%) in semantically unpredictable sentences for the participating systems among all the listeners. Smaller values are better than higher. The performance of the INNOETICS system (letter M) is marked in red.*

5. Discussion

This year's Blizzard set a very interesting challenge. This differed a lot from a typical lab setting where a corpus is carefully selected and recorded; it involved real data from commercial audiobooks produced with its actual target audience in mind rather than the hypotheses and constraints of a research task. Tasks involving audiobooks and storytelling, such as this year's task, indeed offer a very rich ground for research and experimentation and there are still many aspects that have yet to be successfully tackled by TTS systems.

The performance of our system was very high, given the richness of the content and the completely uncontrolled conditions under which it had been recorded. In all the paragraph tasks (evaluating various aspects of the synthetic speech) and the sentence tasks regarding naturalness, our system ranked at the second place, in most cases without any statistically significant difference from the first system.

We expect that an even better performance could be achieved with a larger database. As the content is quite rich in terms of speaking styles and prosodic patterns, we would expect that a larger database would offer better prosodic coverage, i.e. a richer pool of units in different prosodic contexts. Additionally, the fact that the training corpus comprises texts addressed to younger readers and uses a rather limited vocabulary, probably had an impact on the phonetic coverage and balance of the database, and its ability to cope with different

types of texts. This could be problematic, especially when reading text that contains proper names (e.g. news) and less frequent words with more rare phonetic content (e.g. literature for older age groups).

We believe that a strong point of our system is its unit selection module and the fact that it adopts a data-driven approaches where possible (rather than explicit modelling) to address many of the underlying issues involved. Also, the fact that the synthesis process is in tune with the segmentation/analysis process, thus ensuring consistency. Another strong point lies on the robustness of the voice production process.

As typical with concatenative systems, it can be quite challenging to effectively handle cases where diphones units are sparse or unavailable in the database for specific phonetic/prosodic contexts. For such cases, an efficient method resorting to demiphones and/or some type of smoothing to minimize audible artefacts at the concatenation points could improve the speech quality achieved.

In overall, our system achieved great results for a task as challenging as reading highly expressive children audiobooks. This provides evidence that the underlying technology is capable of achieving top quality synthetic speech that can meet the requirements of real world applications in this area.

Highly expressive speech presents a range of new challenges. We have tried out a set of promising ideas but could not fully integrate all of them to our system due to time constraints. The rewarding results we obtained and the lessons we learned in the course, allow us to believe that mimic expressivity is now fully possible. But further to that, convincingly imitating expression in speech is now within reach. We hope that we will be able to verify that in next year's Blizzard Challenge which will involve more data from this domain.

6. Conclusions

The large variability and rich patterns found in expressive speech do not lend themselves easily neither to top-down modeling (i.e. abiding by specific emotion theories and trying to fit their models to the data) nor to bottom-up, unsupervised clustering (i.e. seeking to reveal some underlying hidden structure from the data itself). Various approaches that perform reasonably well in more constrained content seem to break down when faced with the nuances of expressiveness. Many of our attempts to extract some structure from the Blizzard structure did not seem to perform as efficiently as expected, including, for instance, efforts to predict prosodic phrase boundaries or to find significant correlations between prosody and affective word dictionaries. There just seem to be many significant features and forms involved in expressive speech which are not easily (if at all) inferable from the linguistic surface.

An important constituent that still seems to be somehow missing from the expressive/emotional speech synthesis research is, of course, the lack of adequate evaluation methods and protocols for listening tests that could assess the performance of the different systems in relevant dimensions. Standard listening tests do not seem to be able to sufficiently assess the expressiveness/emotion in TTS. New evaluation methods and protocols for listening tests need to be derived. There are quite a few approaches, but most are based on specific emotion taxonomies and pre-annotated databases, which can be controversial and questionable, especially for databases such as this year's children audiobooks. There seems to be a need for a fresh look and a wider discussion in order to build some consensus on that; possibly starting from redefining the task and what we actually expect from expressive/emotional TTS systems, which may indeed differ quite a lot depending on the application.

7. Acknowledgements

We would like to thank the organizers and all the people involved in running the Blizzard Challenge for the time and effort they devote in this highly valued international scientific contest for over a decade.

8. References

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

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

[3] S. Karabetsos, P. Tsiakoulis, A, Chalamandaris and S. Raptis, "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] P. Tsiakoulis, S. Karabetsos, A. Chalamandaris and S, Raptis, "An Overview of the ILSP Unit Selection Text-to-Speech Synthesis System," *Artificial Intelligence: Methods and Applications*, Volume 8445 of the series Lecture Notes in Computer Science, Springer International Publishing, 2014, pp. 370-383

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

[6] A. Chalamandaris, P. Tsiakoulis, S. Karabetsos and S. Raptis, "The ILSP / INNOETICS Text-to-Speech System for the Blizzard Challenge 2014", *Proceedings of the Blizzard Workshop 2014*, http://festvox.org/blizzard/blizzard2014.html

[7] A. Chalamandaris, P. Tsiakoulis, S. Karabetsos, and S. Raptis, "Using Audio Books for Training a Text-to-Speech System", *In Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*, May 26 – 31, Reykjavik, Iceland, 2014, pp. 3076-3080

[8] S. Raptis, "Exploring latent structure in expressive speech", *in Proceedings of the IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)*, December 2-5, Budapest, Hungary, 2013, pp. 741-746

[9] S. Raptis, S. Karabetsos, A. Chalamandaris and P. Tsiakoulis, "A framework towards expressive speech analysis and synthesis with preliminary results", *Journal on Multimodal User Interfaces*, December 2015, Volume 9, Issue 4, pp 387–394

[10] S. Karabetsos, P. Tsiakoulis, A. Chalamandaris and S. Raptis, "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

[11] N. Braunschweiler, M. Gales and S. Buchholz, "Lightly supervised recognition for automatic alignment of large coherent speech recordings," *In Proceedings of the 11th Annual Conference of the International Speech Communication Association*. Curran Associates, Inc., 2010

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

[13] A. Chalamandaris, P. Tsiakoulis, S. Karabetsos and S. Raptis, "An efficient and robust pitch marking algorithm on the speech waveform for TD-PSOLA", *in Proceedings of the 2009 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, Kuala Lumpur, Nov. 18-19, 2009, pp. 397-401