The ADAPT entry to the Blizzard Challenge 2016

João P. Cabral[1], *Christian Saam*[1], *Eva Vanmassenhove*[2], *Stephen Bradley* [1], *Fasih Haider* [1]

[1] Trinity College Dublin, Ireland
[2] Dublin City University, Ireland

cabralj@tcd.ie, saamc@scss.tcd.ie, vanmassenhove.eva@gmail.com
step15hen@gmail.com, haiderf@tcd.ie

Abstract

This paper describes the text-to-speech synthesis system developed for the Blizzard Challenge 2016 by members of the ADAPT centre and colleagues from associated projects. The task was to build a synthetic voice for reading audiobooks to children, from a speech database of audiobooks around 5 hours long. Our entry system is an HMM-based parametric speech synthesizer which was built using a subset of the database (half the total number of the audiobooks of the full dataset). We only used this subset because it was the best quality data we could obtain under the time constraints posed by the Challenges' deadlines. The main parts of the work undertaken on the development of the system for this challenge were on text chunking, including splitting of sentences and segments of text in quotes, and automatic alignment of speech and text data. We also aimed to synthesize speech with emotions to improve the expressiveness of the synthetic speech. Although we could not concretize this task on time for the submission, we plan to carry on this work and possibly use it in a future entry of our system to the Blizzard Challenge.

Index Terms: blizzard challenge, HMM-based synthesis, alignment of speech and text

1. Introduction

In the Blizzard Challenge 2011, a team from the University College Dublin (UCD) that was part of the Centre for Next Generation Localisation (now called the ADAPT Centre) developed the UCD entry system that was based in the unit-selection approach. In contrast, the ADAPT system in this year's challenge is an HMM-based parametric speech synthesizer that was built based on the HTS toolkit [1]. This statistical approach currently has limitations in producing speech that sounds as natural as unit-selection, but it provides a higher degree of parametric flexibility to produce different voice types such as voices with emotions. In this work, a main factor for choosing HMM-based speech synthesis to build our synthetic voice was that we were interested in approximating the expressivity of speech that can be found in audiobooks by modeling the aspect of speech emotions. Another reason was that this approach is fully automatic, whereas unit-selection usually requires manual intervention which can be more time consuming.

In our system a great deal of effort was put on performing high-quality and automatic alignment between the text and speech modalities provided in the database. Another important part of this work was on text analysis for extracting relevant information which could be useful to model some aspects that can contribute to the variability in the voice styles along the audiobooks. We incorporated into our system an emotion prediction tool to synthesize speech with emotions for the dialogue text. However, these tools were not used to synthesize speech for the Blizzard Challenge because we could not build the expressive synthetic voices within the voice building timeline.

In this paper, we start by giving an overview of the HMM-based speech synthesizer. In Section 3 we describe the text processing for chunking the text of the audiobooks into short chunks at the sentence-level that can be used by the synthesizer to build the synthetic voices and synthesize speech for a new audiobook. Next, the process of building the synthetic voices is described in Section 4, focusing on the parts of preparing the speech corpus for building the synthetic voices. The results of the system are discussed in Section 5. Finally, Section 6 gives a summary of the paper and future work.

2. Overview of the System

The ADAPT system is based on the state-of-the-art HTS system available at `http://hts.sp.nitech.ac.jp/`. The block diagram of the HTS system is shown is Figure 1. The system can be divided into the training and synthesis parts. The database consists of sentence-level utterances aligned with the respective text and labels containing the phonetic transcription and other linguistic information. In the training part, the linguistic labels and speech features are used to model context-dependent MSD-HSMMs (Multi-Space Distribution Hidden-Semi Markov Models). The statistical parameters of the MSD-HSMMs are calculated using the maximum likelihood criterion from the phone labels and speech parameters. These models are also clustered using decision trees by using all the contextual factors of the linguistic labels. In the synthesis part, the trained MSD-HSMMs are used to generate the optimal speech parameter sequence to synthesize the input text. Finally, the speech waveform is obtained from the generated parameters using a vocoder.

The text analysis process in the ADAPT system is described in the next section. We used the STRAIGHT vocoder [2] (Matlab version V40_006b) for speech analysis and synthesis, with exception of the Fundamental Frequency (F_0) analysis which was performed using the using the RAPT algorithm implementation of the ESPS tools [3], [4].

3. Text Analysis

The ADAPT system processed the text with the Stanford Core NLP Toolkit [5], which consists of a pipeline of NLP tools including, among many others, a tokenizer, sentence splitter, and POS-tagger. Before doing this however, the system first separated the quoted dialogue from the text using regular expressions. These chunks of dialogue were replaced with tags with

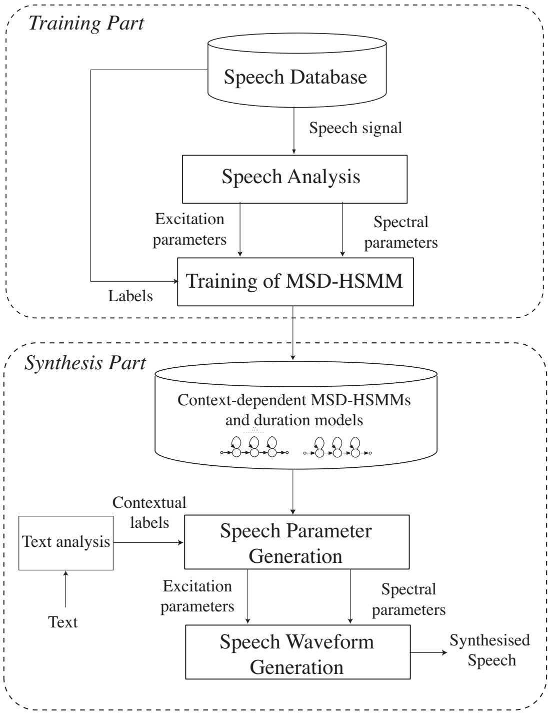

Figure 1: *General block diagram of the HTS system.*

numbered ids. This allowed the dependency parser to recognize entire quotations as subordinates of their speakers (the story characters), or their expressive verbs. This dependency information, as well as part-of-speech and named-entity data was collected using the Stanford Toolkit, but due to time constraints, the component which would make use of the character IDs was not implemented.

Using the Stanford Sentence Splitter, the full text was chunked at sentence boundaries, and then at quotation boundaries. We separated the quoted dialogue from the normal narrative text because we assumed that there was a significant difference in voice style between these two types of text.

Table 1 shows a sample of the chunked output with information about the "type" of text. We categorized text as either normal narrative text (type 0), text before a quotation (type 1), text after a quotation (type 2), or text between two quotations in the same sentence (this text type corresponds to the character ID number). The aim of classifying text as the sentence before or after quotation is to extract information from this text to help in the classification of the character associated with direct speech and emotion of the contiguous quoted text.

The test sentences released for synthesis were also chunked at sentence and quotation boundaries during the text processing stage. This step produced a higher number of text label segments than that of the test dataset so we needed to ensure that the resulting speech signals were concatenated to be consistent with the original file names of the test sentences. Below is an sample of one of the test sentences before chunking, followed by the resulting three chunks with their respective file names.

- AroundTheWorldIn80Days_00002_00033_00096 "My only hope," Fix decided, "is to stop Fogg from leaving India."
- AroundTheWorldIn80Days_00002_00033_00096_0 My only hope,
- AroundTheWorldIn80Days_00002_00033_00096_1 Fix decided,
- AroundTheWorldIn80Days_00002_00033_00096_2 is to stop Fogg from leaving India.

Table 1: *This table shows examples of the text chunking.*

Text Chunk	**Text Type**
Passepartout hit him.	0
Wait!	6
cried Fix.	2
It might have seemed I was against you before-	6
You were!	7
said Passepartout	0
Well, yes,	6
agreed Fix,	3
and I still think Fogg's a thief.	6

The text analysis tools of the FESTIVAL Speech Synthesis System [6] were used to extract linguistic labels used to train the context-dependent MSD-HSMMs with HTS. We used the CMU ARCTIC phoneset (48 phones). The structure of the sentence-level linguistic label is similar to that used in the HTS demo and included phonetic features (2 phones preceding the current phoneme and the 2 following phones) and 53 linguistic features (including information at syllable, word, phrase and utterance levels).

A significant amount of time was spent trying to automatically extract text from the PDF documents of the unlabeled part of the dataset. Several approaches were investigated in this work and *pdftotext* from the open source Poppler PDF rendering libraries (`https://poppler.freedesktop.org/`) was chosen for the PDF to text conversion. We also developed a text processing script that tried to infer pagination from control characters, normalize some of the expressive text layout and filter non-content strings. However, when the text for this subset was made available to the Blizzard participants, our efforts in further improving the text extraction were no longer pursued. Unfortunately, we did not have time to build another synthetic voice with the full dataset that included this additional text.

Recently we proposed a novel emotion labelling system [7] that uses both the information of the emotional polarity from sentiment analysis and emotion category to classify a sentence into one of a set of basic emotions. We intended to use this emotion prediction tool to select the type of voice (one of the emotions or neutral) of the TTS system, given an input sentence. In [7] we also proposed to combine the emotion prediction from text with a speech clustering method to select the utterances with emotion during the process of building the emotional corpus for the speech synthesiser. The plan was to build synthetic voices with emotion by using the HSMM adaptation techniques of HTS. However, the expressive voices were not built on time to be used in the synthesis of the test dataset.

4. Synthetic Voice

4.1. Automatic Alignment Between Text and Speech

The alignment process used in this work was a mixture of guided recognition and forced alignment. The aim was to allow for the automatic alignment over long audio files with potentially unreliable or not fully expanded reference transcripts (e.g. containing abbreviations, numerals etc.).

The general procedure can be divided into the following steps:

1. The audio was partitioned into sentence-like segments based on a phone recognition with a phone language model, no reference was made to the transcripts.
2. The transcripts were processed to provide expansions for numbers, time and date expressions and other text not represented by individual words.
3. The expanded reference text was segmented into sentences or chunks derived from the text chunking method described in Section 3.
4. A 3-gram language model (LM) with Kneser-Ney smoothing [8] was trained on the expanded transcripts.
5. A recognition pass was made on the audio segments with this LM to yield recognition lattices.
6. Two iterations of speaker based unsupervised feature space adaptation (fMLLR [9]) and recognition were performed.
7. The resulting per segment lattices were then concatenated into a super lattice and converted to the word level.
8. The expanded reference text was converted into a word level acceptor finite state transducer (FST).
9. Reference and hypothesis edit distance graphs were created from the super lattice and transcript FST, then they were composed and the shortest path was computed similar to [10].
10. Segments for which the first or last words were not recognized correctly were joined with the preceding or following segment respectively, in order to correct errors that resulted from the initial segmentation.
11. A forced alignment was performed over the joined audio segments.
12. The segments resulting from the alignment were split again between the first and last reference words recovered at a join point.

The steps 1 and 2 of text normalization and expansion were performed using the Festival Speech Synthesis System [6]. In step 4, the LM was trained using NGram library of the OpenGrm Open-source Finite-state Grammar Software Libraries [11]. FSTs were manipulated using the OpenFST libraries [12] in steps 7 to 9. The overall alignment process was realized with the Kaldi Speech Recognition Toolkit [13]. The acoustic models used in the aligner were standard HMM/GMM models discriminatively trained with MMI criterion [14] on the TED-LIUM corpus [15] using the CMU Pronouncing Dictionary [16]. The timing information contained in the labelled part of the released dataset was not used since our goals were to use one process for all files and use our own segmentation.

The word error rate of the hypothesis calculated after the alignment performed at stage 11 was 4.11% for the labelled part of the dataset. From informal analysis, these errors were mainly due to dropping of relevant audio in the first segmentation, but we found that other factors also contributed to the errors such as bad word expansions and some genuine mis-recognition.

4.2. Speech Corpus

The dataset of audiobooks released for the Speech Synthesis Blizzard Challenge 2016 consists of speech and text data of professional audiobooks and includes about 5 hours of British English speech data (sampled at 44 kHz) from a single female talker. It consists of English recordings of John Greenam reading books by Mark Twain (from http://librivox.org/). However, for part of this dataset the tales were provided in PDF format. Initially, we excluded these audiobooks, because we found problems in the extraction of the text from the PDF files and to avoid the effect of any errors in the text extraction. This subset of the database consisted of 25 fairy-tales (half the total number of the audiobooks of the Blizzard dataset).

4.3. Training of the Synthetic Voice

We used the default the training scripts of the HTS system [1] (version 2.3) to build our synthetic voices. In the speech analysis part, the Fast Fourier Transform (FFT) parameters of the envelope and aperiodicity computed using STRAIGHT were converted to 39th order mel-cepstral coefficients, while the FFT parameters of aperiodicity were weighted in 25 frequency bands. Acoustic modeling was performed using the standard five-state left-to-right MSD-HSMM structure and both the state output density function and the state duration were modeled by a single Gaussian distribution in the form of diagonal covariance matrix. The feature vector consisted of five streams: mel-spectrum, aperiodicity parameters, F_0 and its Δ and Δ^2 parameters. The spectrum and aperiodic feature vectors were defined by their static and dynamic parameters (Δ and Δ^2), respectively. Finally, the synthetic voice was built with the HTS toolkit using the speech and text data of the audiobook corpus described in Section 4.2.

5. Evaluation and Results

5.1. Listening Evaluation

The evaluation was conducted online by the Blizzard Challenge organizers and the tasks of the experiment are summarized below:

- Mean opinion scores (MOS), on a scale from 1 to 60, of the different aspects evaluated on book paragraphs: Overall, pleasantness, speech pauses, stress, intonation, emotion, and listening effort.
- MOS (on a scale from 1 to 5) of the speech naturalness of book sentences.
- MOS (on a scale from 1 to 5) of the similarity of the synthesized book sentences to the voice in the reference samples of the original speaker.
- Word error rate (WER), in percentage, of semantically unpredictable sentences (SUS) that listeners heard only once at a time and had to type what they heard.

All the samples were from children's book domain, with the exception of the SUS. The listener types considered in the experiment are: paid participants, volunteers, speech experts and native/non-native listeners. There were 17 systems taking part in the experiment, but only 16 were used in the SUS part as there was no natural speech in this case. These systems can be divided into the group of systems from participants (represented by the letters E to Q) and four reference systems (letters A to D) that are described below:

- System A: Natural speech.
- System B: Unit-selection benchmark system: unit-selection benchmark system (voice built using the same method as used in the CSTR entry to Blizzard 2007).

- System C: HTS benchmark system (voice built using the HTS toolkit).
- System D: DNN benchmark system (voice built using a new toolkit from CSTR, University of Edinburgh).

The organizers of the Blizzard Challenge provided to participants the results, including plots and descriptive statistics: median, MAD (median absolute deviation), mean, standard deviation, n (the number of data points used in the calculations) and na (the number of data points excluded, due to missing data). Additional statistical significance values obtained from Wilcoxon's signed rank tests were also provided to indicate if two systems are statistically different in the different sections and conditions of the experiment.

In the next sections, the box plots for MOS represent the median (central solid bar), quartiles (shaded box), 1.5*quartile range (extended lines) and outliers (circles). Please note that the system orderings in these plots are provided by their mean MOS score, which is speculative. On the contrary, the median scores and statistical difference between the systems can be used to draw conclusions about the MOS results of the systems. The WER results are presented on bar plots of the means. We changed the colors of boxplots presented in the following sections to show the systems that are statistically different from the ADAPT system, which is represented by the letter G and the magenta color.

5.2. MOS for audiobook paragraphs

Figure 2 shows the boxplots of the results obtained in the evaluation of overall impression of the systems when listeners were asked to hear audiobook paragraphs. The results indicate that six systems (excluding natural speech) are significantly better than our system while four systems were significantly worse. The difference between the remaining five systems and our entry are not significant.

It is interesting that the ADAPT system was better than the benchmark HTS regarding the overall impression, because our entry also uses the HTS toolkit for building the HMM-based voice. However, it is difficult to infer about the factors which contributed to this difference because the two systems are not statistically different in terms of the other aspects evaluated in this part of the test (pleasantness, emotion, speech pauses, etc.).

The benchmark unit-selection and DNN systems were significantly better than the ADAPT entry in the paragraph part, with exception of the speech pauses test in which the unit-selection was not statistically different. In general the ADAPT system obtained a similar type of result for the other factors evaluated in audiobook paragraphs, obtaining approximately a middle rank position. The number of systems that obtained significantly better results ranged from six to eight, with exception of emotion evaluation in which our system obtained a place in the bottom of the rank (was significantly worse than 12 other systems). We did not complete our attempt to synthesize speech with emotions so it was expected our system was not significantly different from the benchmark HTS in this task.

The analysis of comparison between the different listener types showed that the MOS for paid listeners were slightly small for all test factors, whereas the MOS for expert listeners were always slightly higher. However, in terms of significant differences between the systems the results are similar for overall impression and in the remaining factors there is a decrease in the number of systems significantly different from the ADAPT system because the number of data points is lower for paid and expert listeners.

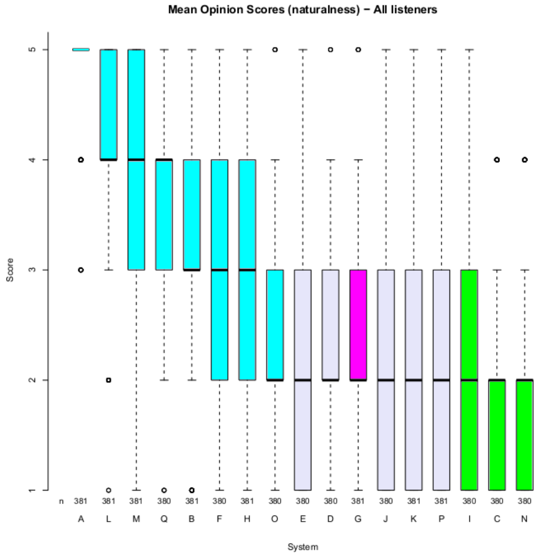

Figure 2: MOS of overall impression obtained in audiobook paragraphs part of the experiment and considering all listeners. The ADAPT system is represented by magenta, while navy blue and green represent the significantly better and worse systems, respectively.

5.3. MOS of naturalness for audiobook sentences

Figure 3 shows the results of speech naturalness considering all listeners. The ADAPT system obtained significantly worse results than seven systems and it was significantly better than three systems, including the benchmark HTS system. The results of the ADAPT system were similar for all listener types (obtained median MOS of 2 and approximately the same number of systems with significantly different scores). In our opinion, the good result obtained against the HTS benchmark is positive because we only used around half the size of the audiobook data available for building our synthetic voice. We expect that this improvement compared with the benchmark is due to better quality of the speech segmentation using our own automatic method for alignment of the audio and text. We also used different text chunking from that provided in the Blizzard dataset. We chunked the text on the sentence level and separated the text in quotes from the "narrator text style", whereas the dataset could include multiple sentences and quoted text within the same text label. This difference in the basic utterance structure could have an effect on the speech prosody modeling at sentence level.

5.4. MOS of similarity for audiobook sentences

Figure 4 shows the results of MOS in terms of speaker similarity considering all listeners. The similarity results of the ADAPT system were not as good as for naturalness because it was significantly worse than ten systems and it was not statistically different from the other five systems (including the benchmark HTS). This result is expected because typically HMM-based speech synthesis obtain poor results in this task due to the limitations of the parametric speech model to reproduce well the voice characteristics of the speaker. For the other listener types in the similarity test, generally the median value o the ADAPT entry was also equal to 2 with exception of paid listeners that obtained median value of 1. However, for these listener types

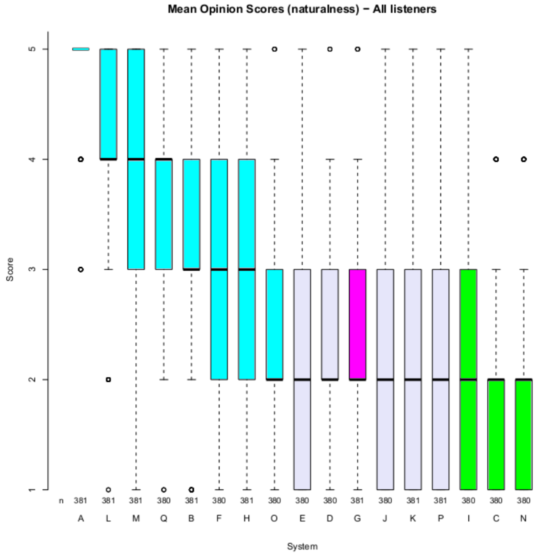

Figure 3: MOS of speech naturalness in audiobook sentences considering all listeners. The ADAPT system is represented by magenta, while navy blue and green represent the significantly better and worse systems, respectively.

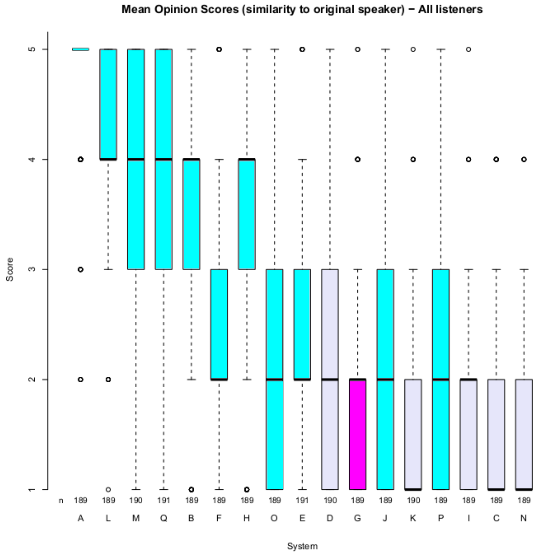

Figure 4: MOS of speaker similarity in audiobook sentences considering all listeners. The ADAPT system is represented by magenta, while navy blue and green represent the significantly better and worse systems, respectively.

the number of systems significantly better than ADAPT were lower compared with the results taking into account all listeners, which is explained by the reduction in the number of data points to compute the scores.

5.5. WER of SUS sentences

In the speech intelligibility test the ADAPT system performed much better than in the naturalness and similarity tests. Figure 5 shows the results of WER obtained by the systems. Our entry system obtained a mean WER of 32%. Only the system L was significantly better than ADAPT with a mean value of 26%. Our system obtained similar mean scores as the benchmark systems, but it was significantly better in terms of intelligibility than seven other systems. We observed a high variation of WER rate between listener types for our system, ranging from 20% for paid listeners to 54% for volunteers (close to the 45% obtained for speech experts). For paid listeners the ADAPT system was significantly better than five systems despite the reduction in the number of data points. But for speech experts there were only two systems significantly different from ADAPT (with higher mean WER values).

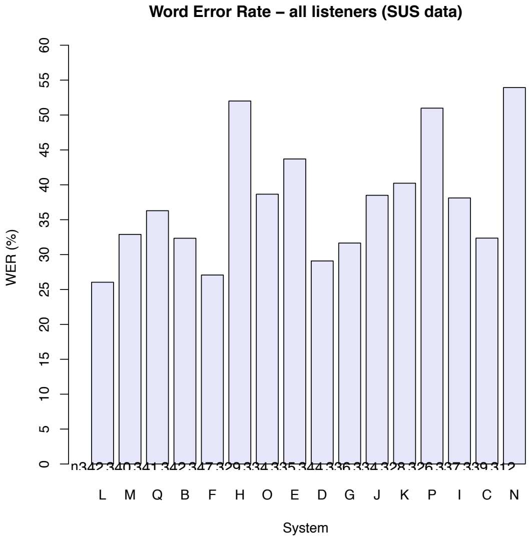

Figure 5: Results of speech intelligibility in audiobook sentences considering all listeners.

6. Conclusions and Future Work

Our ADAPT system for the Blizzard Challenge was an HMM-based Speech Synthesis System. We used the HTS toolkit to build our synthetic voice similarly to the benchmark HTS system. However, we performed our own sentence splitting and alignment between the text and audio, instead of using the text and audio segmentation made available to the participants (for part of the Blizzard dataset). For another part of the dataset the audiobooks were provided only with the full text in PDF format and respective audio. The task of extracting the text from the PDF documents took us significant longer time than we expected and we did not have time to use this part of the data for building the synthetic voice. For us, the results obtained by the ADAPT system were positive because it performed better than the benchmark HTS system at the paragraph level (in terms of overall impression) and sentence level (for speech naturalness), while the two systems were not statistically different for the rest of the tasks. These results were surprisingly good given that we did not use the full dataset for building the synthetic voice. Our explanation for this difference in quality between the two systems is the good quality of the alignment and segmentation of the ADAPT system. We also developed a component for predicting emotions from text with the aim of producing a more expressive synthetic voice. However, we did not have enough time available to complete the development of the synthetic voices with emotions. This explains the poor results of our system in the perception of emotion in the paragraph section of the Blizzard listening test. Although it was disappointing that we did not have enough time to complete all the components that we planned to integrate into the system, we feel encouraged that such components could contribute to improvements of our system in future challenges to synthesize children's audiobooks. More specifically, we plan to improve the performance of our system in terms of speech emotion, prosody aspects (stress, pauses, etc.) and the vocoder component of the synthesizer.

7. Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 13/RC/2106) as part of ADAPT (www.adaptcentre.ie) and EU FP7-METALOGUE project under Grant No. 611073, at Trinity College Dublin, and by the Dublin City University Faculty of Engineering & Computing under the Daniel O'Hare Research Scholarship scheme.

8. References

[1] Tokuda, K., Masuko, T., Miyazaki, N. and Kobayashi, T., “Hidden Markov models based on multi-space probability distribution for pitch pattern modeling”, Proc. ICASSP, pp. 229–231, 1999.

[2] Kawahara, H., Masuda-Katsuse, I. and Cheveigné, A., “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f_0 extraction: Possible role of a repetitive structure in sounds”, Speech Communication, Vol. 27, pp. 187–207, 1999.

[3] Talkin, D., ”Voicing epoch determination with dynamic programming”, J. Acoust. Soc. Amer., 85, Supplement 1, 1989.

[4] Talkin, D. and Rowley, J., “Pitch-Synchronous analysis and synthesis for TTS systems”, Proc. of the ESCA Workshop on Speech Synthesis, C. Benoit, Ed., Imprimerie des Ecureuils, Gieres, France, 1990.

[5] Manning, D. J., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J. and McCloskey, D., “The Stanford CoreNLP Natural Language Processing Toolkit”, Proc. ICASSP, pp. 229–231, 2014.

[6] Black, A. W. , Taylor, P. and Caley, R. “The festival speech synthesis system.” `http://festvox.org/festival/`.

[7] Vanmassenhove, E., Cabral, J. P. and Haider, F., “Prediction of emotions from text using sentiment analysis for expressive speech synthesis”, 9th ISCA Speech Synthesis Workshop, Sunnyvale, USA, Septemper, 2016.

[8] Kneser, R. and Ney, H. “Improved backing-off for M-gram language modeling”, in ICASSP-95, vol. 1, pp. 181–184, 1995.

[9] Gales, M. J. F. , “Maximum likelihood linear transformations for HMM-based speech recognition”, Computer Speech & Language, vol. 12, no. 2, pp. 75–98, Apr. 1998.

[10] Mohri, M. “Edit-distance of weighted automata: General definitions and algorithms”, International Journal of Foundations of Computer Science, vol. 14, no. 6, pp. 957–982, 2003.

[11] Roark, B., Sproat, R., Allauzen, C., Riley, M., Sorensen, J. and Tai, T., “The OpenGrm Open-source Finite-state Grammar Software Libraries”, in Proc. of the ACL 2012 System Demonstrations, Stroudsburg, PA, USA, pp. 61–66, 2012.

[12] Allauzen, C., Riley, M., Schalkwyk, J., Skut, W. and Mohri, M. “OpenFst: A General and Efficient Weighted Finite-State Transducer Library”, in Proceedings of the Ninth International Conference on Implementation and Application of Automata (CIAA 2007), vol. 4783, pp. 11–23, 2007.

[13] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G. and Vesely, K. “The Kaldi Speech Recognition Toolkit”, in IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Hilton Waikoloa Village, Big Island, Hawaii, US, 2011.

[14] Bahl, L., Brown, P., de Souza, P., and Mercer, R. “Maximum mutual information estimation of hidden Markov model parameters for speech recognition”, in Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP ’86., vol. 11, pp. 49–52, 1986.

[15] Rousseau, A., Deléglise, P. and Estève, Y. “TED-LIUM: an Automatic Speech Recognition dedicated corpus”, in Proc. of the Eight International Conference on Language Resources and Evaluation (LREC’12), Istanbul, Turkey, pp. 125–129, 2012.

[16] Rudnicky, A., “The Carnegie Mellon Pronouncing Dictionary.” `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`.