SYNTHESIZING EXPRESSIVE SPEECH FROM AMATEUR AUDIOBOOK RECORDINGS

Éva Székely[1], *Tamás Gábor Csapó*[2], *Bálint Tóth*[2], *Péter Mihajlik*[2,3], *Julie Carson-Berndsen*[1]

[1]CNGL, School of Computer Science and Informatics, University College Dublin, Dublin, Ireland
[2]Department of Telecommunication and Media Informatics,
Budapest University of Technology and Economics, Budapest, Hungary
[3]THINKTech Research Center, Vác, Hungary

eva.szekely@ucdconnect.ie, {csapot, toth.b, mihajlik}@tmit.bme.hu, julie.berndsen@ucd.ie

ABSTRACT

Freely available audiobooks are a rich resource of expressive speech recordings that can be used for the purposes of speech synthesis. Natural sounding, expressive synthetic voices have previously been built from audiobooks that contained large amounts of highly expressive speech recorded from a professionally trained speaker. The majority of freely available audiobooks, however, are read by amateur speakers, are shorter and contain less expressive (less emphatic, less emotional, etc.) speech both in terms of quality and quantity. Synthesizing expressive speech from a typical online audiobook therefore poses many challenges. In this work we address these challenges by applying a method consisting of minimally supervised techniques to align the text with the recorded speech, select groups of expressive speech segments and build expressive voices for hidden Markov-model based synthesis using speaker adaptation. Subjective listening tests have shown that the expressive synthetic speech generated with this method is often able to produce utterances suited to an emotional message. We used a restricted amount of speech data in our experiment, in order to show that the method is generally applicable to most typical audiobooks widely available online.

Index Terms— Speech synthesis, expressive speech, audiobook, language resources

1. INTRODUCTION

Recording speech corpora specifically designed for expressive speech synthesis requires significant time and resources [1]. Being able to leverage pre-existing language resources containing expressive speech such as public domain audiobooks would significantly increase the range of available expressive synthetic voices, making it possible to build synthetic voices from speakers of different gender, age, accent and voice characteristics. The availability of this variety would widen the application area of expressive speech synthesis, as the voices would meet individual needs in applications such as speech generating devices (SGD) for non-speaking individuals, speech-to-speech translation applications and dialogue systems using intelligent agents. For a language in which there are limited language resources such as Hungarian, this could make a significant contribution.

Previous approaches to synthesizing expressive speech in Hungarian applied emotion-specific speech inventories uttered by a professional speaker for use in a diphone/triphone TTS (Text-To-Speech) prototype [2]. In the work presented in this paper we investigate the use of public domain audiobooks for expressive speech synthesis; in particular we aim at the most widely available type of public domain audiobooks: those recorded by non-professional speakers.

A number of studies have previously dealt with synthetic voices from audiobooks. [3] explores the use of an audiobook where a storyteller performs multiple roles. An expressive unit selection TTS is created and it is shown that listeners usually recognize the neutral, young, elder and adult voice styles. [4] extends the traditional forced alignment to large speech files and uses this for HMM (Hidden Markov-Model) based speech synthesis. [5] introduces lightly supervised recognition for speech-text alignment and applies this for a number of audiobooks. [6] follows this direction and improves the expressiveness of statistical speech synthesis systems. Building expressive synthetic voices from audiobooks generally involves three basic steps:

1. The audio which is recorded in large sequences needs to be aligned with the corresponding text.
2. A variety of expressive voice styles need to be identified and grouped together.
3. The synthetic voices can be built.

The studies mentioned above use large size audiobooks containing a high proportion of expressive speech, that are often

This research is partially supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at University College Dublin (UCD). The opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Science Foundation Ireland. This research is further supported by the CESAR project (Grant No 271022) and by the TÁMOP-4.2.1/B-09/1/KMR-2010-0002 project. The authors would like to thank Ingmar Steiner for the valuable comments that have improved the manuscript.

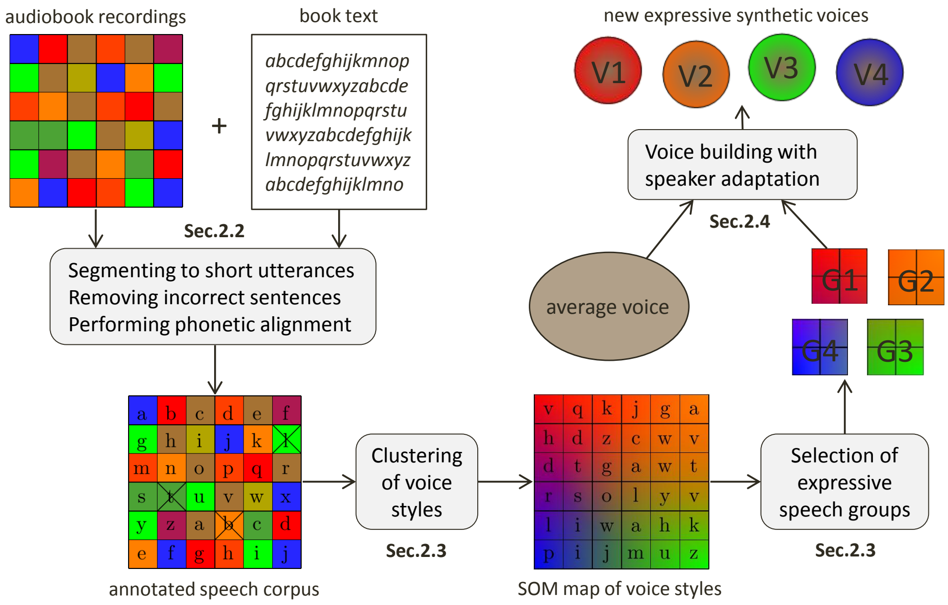

Fig. 1. Overview of the proposed system. The colors indicate different voice styles present in the audiobook. Sections 2.2-2.4 describe the different steps in detail.

read by exceptionally talented speakers. Using more commonly available audiobooks for speech synthesis, recorded and read by amateur speakers, poses (at least) three additional challenges:

- The recording quality may not be consistent throughout the audiobook.
- The amount of mismatches between speech and text may be significant due to misreadings.
- The proportion of expressive speech may be significantly smaller than the proportion of neutral "read speech" in the recordings.

This often results in the amount of speech data usable for speech synthesis being considerably reduced. In the next sections, we present the steps we have taken to address the above challenges. The aim of this paper is to develop an end-to-end minimally supervised method for expressive voice building from audiobooks that can be successfully applied to most open source audiobooks of average length and quality.

2. DATABASE AND METHODS

Fig. 1 shows the overview of our proposed method. First, speech recognition is applied on the audiobook, which results in speech segments with the corresponding aligned text. These segments are the input of unsupervised expressive clustering yielding several subsets differing in their style. The subsets are fed into a statistical speech synthesis system resulting in expressive voices. The next paragraphs present the building blocks in detail.

2.1. Database

The speech corpus used in this study is a Hungarian audiobook read by a female non-professional voice actor. The audiobook *Egri csillagok* (*Eclipse of the Crescent Moon*) by Géza Gárdonyi was downloaded from LibriVox[1], an on-line community of volunteer readers making available recordings of public domain (out-of-copyright) books in numerous languages; the audio recordings on Librivox are in the public domain as well, with no restriction on usage. The selected audiobook was read by a young female amateur speaker in quiet conditions with a low-end microphone. The 44 kHz, 128 kbps MP3 format of the speech material was converted to uncompressed PCM format and downsampled to 16 kHz for further use. The audiobook consists of 17.5 hours of speech recordings in 92 chapters, each chapter consisting of one audio file with the corresponding text. The text contains 136k running words and 918k running letters. Altogether 26k different word forms were found. In this study, only the first 18 chapters (141 minutes) were used, as the recording conditions were found to be inconsistent in the later chapters. Another reason for using this subset is that audiobooks of similar length are widely accessible, so the method is more generally applicable.

This audiobook poses many challenges for the speech-text alignment task, voice style separation task, as well as for the voice building task. The book read by the speaker was written in the 19th century, and the story is set in the 16th century, so the corpus contains a great number of uncommon words. As the book contains a lot of narration and few dialogues, the speaker rarely gets an opportunity to use her voice in expres-

[1] http://www.librivox.org

sive ways. This results in small amounts and less intensive expressive speech, not enough to build an individual synthetic voice, but quantitatively enough for speaker adaptation.

2.2. Corpus segmentation and alignment

The preparation of the speech corpus for TTS, involving automatic speech recognition, consists of three main tasks:

1. Segmenting long audio files into "style-homogeneous" short utterances.
2. Removing incorrectly read sentences.
3. Performing phonetic alignment on the remaining utterances.

Based on the idea of [5] a large vocabulary continuous speech recognition (LVCSR) system is used to align the audiobook. The language model component of the LVCSR system [7] was trained on the whole original text of the novel. A word trigram model was trained using the SRILM Toolkit [8] applying Good-Turing discounting.

A full pronunciation lexicon was used and phonemic pronunciations were generated automatically by applying grapheme-to-phoneme rules [7]. Acoustic models were speaker independent, cross-word, tied-state triphones trained in the classical Maximum Likelihood way on the MRBA database [9][2]. The acoustic model training was performed with HTK [10] resulting in $\sim$2000 shared states, where at most seven Gaussian mixtures were applied per state. Standard MFCC features were used with first and second time derivatives and blind channel equalization was also applied to the 8 kHz bandwidth data. VOXerver [7] WFST-decoder was used for automatic transcription. The decoder has built in capabilities for HMM-based speech silence detection and indexing based on HMM state duration. The authors used this feature to segment the several-minutes-long chapters into short utterances. Average length of the segmented utterances is about 3 seconds, which means that one written sentence was cut into 3 speech chunks on average.

For categorization of speech, it is important for the utterances to be as homogeneous as possible – and at the same time they should be matched precisely to the corresponding text sequence. To solve this issue, we applied the following technique: punctuation marks in the original text of a given chapter were replaced by special boundary marks. Automatic transcripts of the utterances of a given chapter were reconcatenated, inserting the special boundary marks between utterance texts. These two long strings were matched using dynamic programming alignments as in ASR (Automatic Speech Recognition) error evaluations. All the utterances that had greater Levenshtein-distance than a threshold (2 characters in our case) were excluded from further processing. As a result, $\sim$74% of the speech data were kept and $\sim$52% of misreadings were filtered out. The word mispronunciation rate of the speaker was $\sim$7%, which means a high (over $\sim$80%) false rejection rate. The results may seem inappropriate for direct use in TTS; however, we must emphasize that our primary aim was segmentation to text-aligned short (style-homogenous) utterances, not strict misreading detection. In a previous study, we found that for speech style adaption in HMM-based speech synthesis, time-homogeneous adaptation speech data – in terms of speech style or expression – can be more crucial than immaculate transcription [11]. Phonetic alignment was performed automatically using well-known forced alignment techniques, without manual correction[3].

2.3. Clustering of voice styles

The term *voice style* is used to describe the different ways in which a speaker produces an utterance in terms of changes in voice quality, combined with certain prosodic variation over the course of the entire utterance. The voice styles occurring in audiobooks are not only direct expressions of emotion and affect, but often a result of the speaker changing their voice to portray different characters or to arouse the reader's interest. The voice style separation in this work follows a technique presented in [12], where voice quality parameters of the glottal source are used to identify the variety of speaking styles in an audiobook, and place similar utterances on a continuum of neighboring clusters of a Self-Organizing Feature Map [13]. The unsupervised method has the advantage of being able to use the differences of voice styles within one speaker's corpus, without any previous knowledge or explicit labeling. Previous subjective evaluation [12] showed that this method successfully separated speech data into groups of utterances associated with different voice characteristics and styles.

The clustering method uses glottal source parameters of the Liljencrants-Fant model [14]. The glottal source parameters, open quotient (OQ), return quotient (RQ) and speed quotient (SQ), have been extracted as described in [15]. OQ, RQ, SQ and fundamental frequency were used to perform unsupervised clustering on the speech data with a Self-Organizing Feature Map [13]. Seven input features were calculated for each speech segment: mean and delta values of OQ, RQ, SQ, and average f0 over that speech segment.

This method has been applied for expressive speech synthesis [16], and an adapted version used shimmer and jitter features [6]. Both studies used an audiobook corpus from a middle aged American male who is a professionally trained speaker and uses a great variety of expressive speaking styles. In the current study, the method is modified for use with a more typical example of an open source audiobook, which contains a lower percentage of expressive speech. Because this audiobook was segmented into short utterances during the speech-text alignment phase (Sec. 2.2), no further segmentation is needed to apply the voice style separation method successfully. Within a short speech segment (1-6

[2] http://alpha.tmit.bme.hu/speech/hdbMRBA.php

[3] Alignments of open source audiobooks conducted with this method can be found at: http://thinktech.hu/public/lang-res/HunAuB-ECM-text/

seconds), we can assume that there are no significant voice style changes. Altogether 25 clusters were used to group the similar-sounding utterances. On this speech data, only an informal evaluation by an expert listener was carried out, confirming that the variety of voice styles was successfully mapped. As expected, due to the large narrative portion in the audiobook, expressive voice styles formed smaller clusters.

Based on the topographical information provided by the feature map, four groups of clusters (G1–G4) containing an approximately equal amount of speech recordings ($\sim$10 minutes each) that were at the greatest distance from each other were selected to serve as adaptation corpora in the voice building.

2.4. Voice building

Hidden Markov-model based TTS has the ability to create new voices from rather short speech corpora with speaker adaptation. As few as five sentences are enough to create new voice characteristics that are similar to the original speaker [17]; however to achieve the quality of speaker-dependent HMM-based TTS, longer adaptation corpora are necessary. ASR transcription-based unsupervised adaptation was previously investigated [18] with promising results, even in the case of higher phoneme error rates.

For the average voice, five speech databases were used: four males and one female speech corpora (altogether about 12 hours of speech [11]). The utterances in the average voice database were well-designed, phonetically balanced sentences. The content of the utterances was manually verified. Phone boundaries were determined by forced alignment.

For speaker adaptation, the four groups (G1–G4) of different expressive voice styles (Sec. 2.3) were used, consisting of speech segments and their phonetic alignment (Sec. 2.2). The average voice of the modified Hungarian version of HTS [11] was adapted using CMLLR[17]. Sample sentences were generated with STRAIGHT [19]. For the sake of clarity the four voices built from the cluster groups G1–G4 are referred to as V1–V4, respectively.

3. EVALUATION

Evaluating expressive synthetic speech is not a straightforward task. Besides evaluating *naturalness*, it is desirable to assess whether there are significant *audible differences* among the expressive styles, as well as to evaluate the *appropriateness* of the voice styles for specific tasks or utterances [16, 20]. Because the voice style separation in the corpus was unsupervised, it was not possible to know ahead of the evaluation, in what way the resulting synthetic voices were going to sound different. However, it is necessary to show that the perceived difference between the voices is consistent across utterances and that the differences are functional, i.e., that they represent different expressive voice styles, and are not merely a result of the voice being trained on a different set of sentences which naturally contain slightly different prosodic variation.

3.1. Set-up of the perception test

A subjective web-based listening test was conducted, using 20 sentences synthesized with the four speaker-adapted voices. Test sentences were chosen that hold emotional content of various sentiment and intensity. During a pre-test, it was found that V1 and V2 did not show a significant audible difference. Therefore, in the evaluation only V1, V3, and V4 were used.

The evaluation consisted of two consecutive tasks. The first part of the test was a paired comparison, containing all combinations of the 20 sentences with the three voice styles (60 pairs altogether). Listeners were asked to rate on a scale of 1–5 how much difference they heard between the samples in terms of *voice characteristics and prosody*. In the second part, the three versions of the 20 sentences were presented separately, in randomized order. Subjects had to rate how well (on a scale of 1–5) the voice style characteristics of the speaker suited the *content* of the current utterance. The goal of this suitability task in the evaluation was to assess the usefulness of the expressive features of the synthetic voices in conveying the paralinguistic content of a particular message.

3.2. Results of the perception test

The listening test was completed by 18 participants. The average age of the subjects was 36 years (ranging from 23 to 55). All participants were native speakers of Hungarian with no known hearing impairment.

The results of the discrimination task show that the listeners generally perceive a difference between the sentence variants and that there is a trend that the perceived difference increases in the V1 vs. V3 $<$ V1 vs. V4 $<$ V3 vs. V4 direction. The visualization of the measured distances between the voices is displayed in Fig 2.

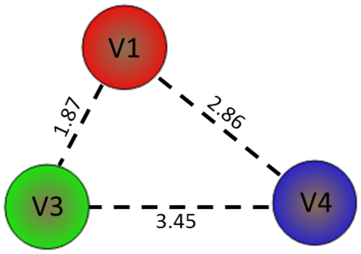

Fig. 2. Result of the discrimination task. The distance between the nodes represent the perceived difference between the voices. (Average scores of the pairwise comparison with 20 sentences.)

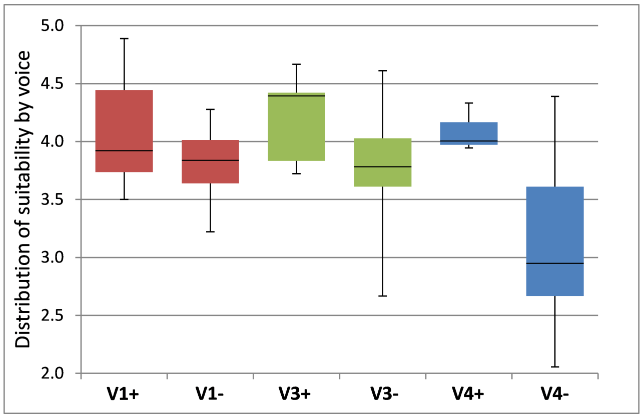

Fig. 3. Results of the suitability task. The y axis shows the overall scores for the three voices: for each voice, the left boxplot shows the scores where the voice was preferred, on the right the scores the respective voice got in cases where another voice was preferred for a particular sentence.

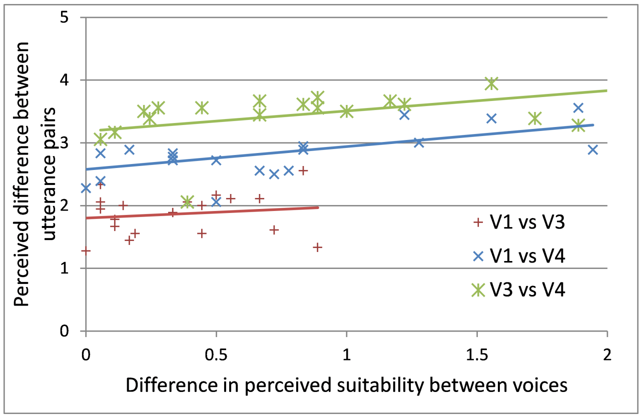

Fig. 4. Scatterplots of the perceived difference between utterance pairs of different synthetic voices presented (y axis) and the difference in perceived suitability of the synthesized utterances when presented individually in the evaluation (x axis)

The overall scores of the suitability task can be seen in Fig. 3, which shows that the voices V1 and V3 received higher scores than V4. Fig. 4 shows the perceived difference between utterance pairs of different synthetic voices presented (first part of the test) and the difference in perceived suitability of the synthesized utterances when presented individually in the evaluation (second part of the test). The differences between the voices are confirmed when the two evaluation results are compared with each other: a larger difference in the first experiment (direct comparison) strongly correlates with a higher difference in perceived suitability in the second test. This is an indication that the perceived differences are due to a consistent difference between the expressive voice styles each voice represents, and not to dissimilarities caused by using a different set of sentences for voice building.

Table 1 shows the group which received the highest mean score in the suitability task for each utterance. The average of the scores of utterances with the best perceived voice style is 4.1. It can be observed that the utterances have different preferred variants supporting our hypothesis that the best suitable voice style depends on the textual content of the sentence.

These results are indicators of each voice style being of added value in an application, as each of the three is the most suitable in a number of sentences presented in the evaluation (V1 and V3 in 8 cases each, V4 in 4 cases was found most suitable).

3.3. Discussion

During the minimally supervised clustering of expressive speech, four cluster groups were created with the hypothesis that each of them represents a different voice style. However, during our pre-test we found that sentences synthesized with V1 and V2 showed only insignificant differences, which may have been caused by the small size of the adaptation database. Another reason might be that the cluster groups G1 and G2 featured subtle voice style differences which were obscured by the nature of the statistical parametric speech synthesis method.

4. CONCLUSION AND FUTURE WORK

The aim of this study was to develop and test an end-to-end minimally supervised expressive voice building method that can be successfully applied to most open source audiobooks of average-length and quality read by amateur speakers. The evaluation has shown that three of the four voices built from the Hungarian audiobook corpus show significant and consistent differences in voice style. Each of these synthetic voices was rated in terms of expressive speaking style as best suitable for several utterances by the listeners in the perception test. This indicates that each voice style could be functionally used in an application needing expressive speech synthesis. With the help of the method presented above, many available audiobooks become possible candidates for the building of a variety of expressive synthetic voices. Utilizing this largely untapped source of speech data could provide significant support for speech applications in under-resourced languages.

Future work involves further testing of the usability of expressive synthetic speech created with this method in applications such a SGDs, book reading systems and speech translation applications.

Table 1. Preferred voice styles of the synthesized utterances

Utterance no	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Preferred voice	V3	V4	V1	V3	V3	V3	V1	V3	V1	V1	V4	V1	V3	V4	V1	V3	V1	V3	V4	V1
Average score	3.8	3.9	3.5	3.9	3.7	4.4	4.8	4.4	4.0	3.8	4.0	4.9	4.4	4.3	4.3	4.5	3.8	3.8	4.7	3.6

5. REFERENCES

[1] N. Campbell, "Databases of emotional speech," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[2] Cs. Zainkó, M. Fék, and G. Németh, "Expressive speech synthesis using emotion-specific speech inventories," *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*, pp. 225–234, 2008.

[3] Y. Zhao, D. Peng, L. Wang, M. Chu, Y. Chen, P. Yu, and J. Guo, "Constructing stylistic synthesis databases from audio books," in *Proc. of Interspeech, Pittsburgh*, 2006, pp. 1750–1753.

[4] K. Prahallad and A.W. Black, "Handling large audio files in audio books for building synthetic voices," in *Proc. of SSW7*, 2010, pp. 148–153.

[5] N. Braunschweiler, M. J. F. Gales, and S. Buchholz, "Lightly supervised recognition for automatic alignment of large coherent speech recordings," in *Proc. of Interspeech, Makuhari*, 2010, pp. 2222–2225.

[6] F. Eyben, S. Buchholz, N. Braunschweiler, J. Latorre, V. Wan, M. J. F. Gales, and K. Knill, "Unsupervised clustering of emotion and voice styles for expressive TTS," in *ICASSP*, 2012.

[7] P. Mihajlik, Z. Tüske, B. Tarján, B. Németh, and T. Fegyó, "Improved recognition of spontaneous Hungarian speech - morphological and acoustic modeling techniques for a less resourced task," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 6, pp. 1588–1600, 2010.

[8] A. Stolcke, "SRILM-an extensible language modeling toolkit," in *Seventh International Conference on Spoken Language Processing*, 2002, pp. 901–904.

[9] K. Vicsi, A. Kocsor, Cs. Teleki, and L. Tóth, "Speech database at a computer using environment," in *II. Magyar Számítógépes Nyelvészeti Konferencia, Szeged*. Szegedi Tudományegyetem Informatikai Tanszékcsoport, 2004, pp. 315–318.

[10] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK book*, 2002.

[11] B. Tóth and G. Németh, "Improvements of Hungarian hidden Markov model-based text-to-speech synthesis," *Acta Cybernetica*, vol. 19, no. 4, pp. 715–731, 2010.

[12] É. Székely, J. P. Cabral, P. Cahill, and J. Carson-Berndsen, "Clustering expressive speech styles in audiobooks using glottal source parameters," in *Proc. of Interspeech, Florence*. 2011, pp. 2409–2412, ISCA.

[13] T. Kohonen, S. Kaski, and H. Lappalainen, "Self-organized formation of various invariant-feature filters in the adaptive-subspace SOM," *Neural Computation*, vol. 9, no. 6, pp. 1321–1344, 1997.

[14] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 26, no. 4, pp. 1–13, 1985.

[15] J. P. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "Towards an improved modeling of the glottal source in statistical parametric speech synthesis," in *Proc. of SSW6*, 2007.

[16] É. Székely, J. P. Cabral, M. Abou-Zleikha, P. Cahill, and J. Carson-Berndsen, "Evaluating expressive speech synthesis from audiobook corpora for conversational phrases," in *Proc. of LREC, Istanbul*, 2012.

[17] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Speaker adaptation for HMM-based speech synthesis system using MLLR," in *Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, 1998.

[18] B. Tóth, T. Fegyó, and G. Németh, "The effects of phoneme errors in speaker adaptation for HMM speech synthesis," in *Proc. of Interspeech, Florence*, 2011, pp. 2805–2808.

[19] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.

[20] E. Eide, A. Aaron, R. Bakis, W. Hamza, M. Picheny, and J. Pitrelli, "A corpus-based approach to expressive speech synthesis," in *Proc. of SSW5*, 2004, pp. 79–84.