Objective measures to improve the selection of training speakers in HMM-based child speech synthesis

Avashna Govender & Febe de Wet
Human Language Technology Research Group, Meraka Institute
Council for Scientific and Industrial Research, Pretoria, South Africa
Email: agovender1@csir.co.za, fdwet@csir.co.za

***Abstract*—Building synthetic child voices is considered a difficult task due to the challenges associated with data collection. As a result, speaker adaptation in conjunction with Hidden Markov Model (HMM)-based synthesis has become prevalent in this domain because the approach caters for limited amounts of data. An initial average voice model is trained using data from multiple speakers and adapted to resemble a specific target child speaker. Due to the scarcity of child speech data, initial models used in this approach are mostly trained with adult speech data. However, selection of appropriate training speakers from large corpora is not a trivial task because there is no means, other than conducting exhaustive subjective listening tests, to determine which training speakers will yield the best quality synthetic child voice. Therefore, there is a need to find an objective measure that can be used to easily identify a small set of training speakers that will yield the best quality output. In this paper we investigate whether a relationship exists between objective and subjective voice evaluation measures with regard to the selection of training speakers for an average voice model used in speaker-adaptive HMM child speech synthesis. Results indicate that, if training speakers that are closer to the target speaker are used to train initial models, better quality child voices are generated.**

I. Introduction

The main objective of research in the field of Text-to-Speech (TTS) synthesis is to generate speech that is as natural and intelligible as that of a human speaker. Concatenative-based synthesis systems have been successful in generating high quality synthetic speech [1]. However, a crucial limitation of this technique is that each unique voice requires a unique set of recordings to be made. Various attempts have been made to generate high quality speech that contains a variety of speakers, voice characteristics and speaking styles [2]. However, to implement similar concatenative-based synthesis systems requires large amounts of speech data from multiple speakers and this data collection process has high costs associated with it.

In contrast to concatenative-based speech synthesis, a statistical parametric speech synthesis system based on hidden Markov models (HMMs) can generate synthetic speech without requiring large scale speech corpora [3]. Such a system has the advantage of easily transforming its models such that a system can reproduce varying speakers, speaking styles and emotions [4]. This process is called speaker adaptation. Speaker adaptation applies transformation techniques that only require small amounts of adaptation data to adapt an already trained system to a specific target speaker [5]. In this way, thousands of voices can be generated, representing a diverse set of speakers and voice characteristics [6].

In the domain of child speech synthesis speaker adaptive synthesis is prevalent because data collection is a major challenge. Speaker adaptation is used in conjunction with HMM-based synthesis to develop child voices. The average-voice based synthesis technique introduced in [7] and applied in [8] uses an average voice model and model adaptation to adapt an initial model, trained with multiple speakers, to resemble a target child speaker. Due to the scarcity of child speech data, sufficient data is usually not available to train an initial average child model. As a result, the initial models used in child speech synthesis are often trained with adult speech data [8], [9].

A drawback of this approach is the method that is used to select training speakers. Identifying the most suitable adult training speakers from large corpora is a challenging task because there is no means, other than conducting exhaustive subjective listening tests, to determine which training speakers would produce the best quality synthetic child voice. Therefore, there is a need to find an objective measure that can be used to easily identify a small set of training speakers that will yield the best quality output.

The aim of this study was to determine whether a relationship exists between objective and subjective voice evaluation measures with regards to the selection of training speakers for an average voice model used in child speaker-adaptive HMM-based synthesis. The hypothesis is that, if training speakers that are *closer* to the target speaker are used to train the initial models, better quality child voices can be generated [10]. To determine exactly what is meant by *closeness* it needs to be defined in terms of objective measures that directly correlate with yielding better quality voices.

This paper is organized as follows. Section II provides a review of previous work and related research. Section III describes the methods used in this study, followed by the

978-1-5090-3334-5/16/$31.00

presentation of results. The results are discussed in Section IV and conclusions are presented in Section V.

II. Background

A. Overview of average-voice-based speech synthesis

HMM-based synthesis is a statistical parametric approach that generates speech waveforms using hidden Markov models [11]. HMMs were originally applied in automatic speech recognition but are also being applied in speech synthesis [12], [13]. HMM-based synthesis has grown in popularity over the last few years, with success in synthesizing both natural sounding and intelligible speech. In HMM-based synthesis systems, human speech production is modeled in terms of the frequency spectrum (vocal tract) and the fundamental frequency (vocal source). In this way, models are trained and not memorized as in the concatenative-based synthesis approach.

Model adaptation is used in conjunction with HMM-based synthesis to implement speaker adaptation [14]. Typically, an initial model is trained with multiple speakers. The initial model is also referred to as an average voice model. The average voice model is adapted using speaker adaptation techniques to transform the initial model in such a way that its properties resemble those of a specific target speaker. This approach is referred to as average-voice based synthesis [7].

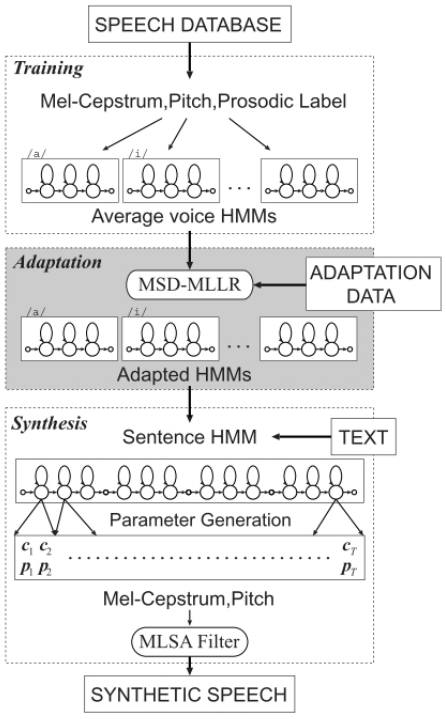

Fig. 1: Schematic representation of an average-voice based synthesis system adapted from [5]

Figure 1 illustrates an average-voice based HMM synthesis system. The architecture of the average-voice based HMM-based speech synthesis system comprises of a training stage, adaptation stage and synthesis stage.

In the training stage, speech analysis is performed on the raw training data. Two important parameters are extracted: the spectral and excitation parameters. First, the static spectral features, mel-cepstral coefficients, are obtained by mel-cepstral analysis [15] and then the excitation parameters, which are the F_0 values, are estimated using an instantaneous frequency amplitude spectrum (IFAS) based method [16]. In addition, dynamic and acceleration features are also used. These are calculated as the first and second order regression coefficients of the static features. The features extracted from the speech database are modeled using multi-stream HMMs. The spectral features are modeled using continuous probability distributions and the excitation features are modeled using multi-space probability distributions (MSD). In order to model variations of spectrum, pitch and duration, phonetic and linguistic contextual factors and stress related factors need to be taken into account.

Using context-dependent HMMs, an average voice model is trained as the initial model using training data from each speaker in a multi-speaker speech database. A decision-tree-based context clustering technique is subsequently applied separately to the spectral and pitch parts of the average voice HMMs. Since it is impossible to prepare training data which cover all possible context-dependent units, a context clustering technique needs to be applied. This technique is used to cluster HMM states and share model parameters among states in each cluster. State duration distributions of the average voice model are obtained in the same way by applying the same clustering technique [4].

The average model is used to bootstrap the speaker adaptation process. The average voice model plays a crucial role in the adaptation process whereby the voice characteristics and fundamental frequency of the average voice model are simultaneously transformed into that of a target speaker using only a small amount of speech data uttered by the target speaker (adaptation data) [7]. This transformation is performed by first calculating feature vectors from the adaptation data. The average voice HMMs are then transformed into the target speaker HMMs by applying a speaker adaptation technique.

MLLR (Maximum likelihood linear regression) adaptation is usually applied for the adaptation of spectral and F_0 features. In practice, several linear regression functions can be applied and are typically derived from the MLLR and maximum a posteriori (MAP) algorithms. Constrained Structural Maximum a Posteriori Linear Regression (CSMAPLR) is one of the most recently proposed approaches and has proven to generate the best quality adapted synthetic speech [17].

Finally, during the synthesis part, HMMs are concatenated according to the arbitrary input text given to the synthesiser. First, a phoneme sequence is constructed by the front-end. Then an utterance HMM is constructed by concatenating the relevant context-dependent HMMs. The state durations of the utterance HMM are then determined based on state duration densities. A parameter generation algorithm subsequently generates the sequence of spectral and excitation parameters that maximize their output probabilities. Usually,

978-1-5090-3334-5/16/$31.00 ©2016 IEEE

a maximum likelihood (ML) criterion is used to estimate the model parameters [13]. Lastly, a speech synthesis filter such as the mel log spectrum approximation (MLSA) filter is used to synthesise the speech waveform using the generated parameters [18].

In [17], it was shown that using the average-voice speech synthesis approach, natural speech could be obtained for an adult target speaker using as little as 100 utterances of adaptation data. This corresponds to approximately six minutes of speech data. In addition, the synthetic voices generated by adapted models were compared with synthetic speech generated by a conventional, speaker dependent HMM-based speech synthesis system and their results showed that using average-voice based synthesis produced more natural sounding synthetic speech than the speech produced by a speaker dependent system.

This observation was said to be the result of a data rich average voice model that provides strong prior knowledge for speech generation with the adaptation data being used to estimate speaker specific characteristics. The average voice model utilizes a large variety of contextual information included in the multi-speaker database as a priori information for speaker adaptation and therefore provides a robust basis for synthesizing speech for a new target speaker. As a result, synthetic speech of the target speaker can be obtained robustly even if only a limited number of speech samples are available for the target speaker.

B. HMM adaptation applied to child speech synthesis

One of the major challenges in the development of child speech synthesis systems is data collection. Finding children who are willing and/or able to record many hours of useful data is one of the first problems faced. If a suitable candidate can be found, the speech data that can be obtained is usually too little and invariably contains imperfections that are not suitable to synthesise a high quality voice [8]. Therefore, in order to synthesise child speech, an approach that can handle a limited amount of speech data is required. Fortunately, the adaptation ability of HMM-based synthesis provides a promising solution. HMM adaptation has been widely used to synthesise adult voices but very little work has been done to create children's voices using this technique.

Even though HMM adaptation is a viable solution, the scarcity of child speech data still poses a challenge to speech synthesis using HMM adaptation because the available data is usually insufficient to train an initial average child model. As a result, initial models in child speech synthesis are most often trained with adult speech data [8], [9]. Using this approach, child voices have been successfully synthesised. However, the quality of these voices are not suitable for commercial or real-world applications.

In [9], it was shown that using a gender-independent initial model yielded better results for child speech synthesis than using gender-dependent initial models. This observation is in contrast with results that were obtained in studies on adult speech synthesis where gender-dependent initial models were found to perform better than their gender-independent counterparts[17].

In [10], it was observed that the distance between the average voice model and the target speaker can affect the quality of the resulting adapted voices. This result implies that the use of an adult average voice model for adaptation to a child target speaker in itself will result in poor results. However, if training data is only selected from the adult speakers that are closest to the target speaker, then the overall voice quality could be improved.

A correlation between naturalness and *closeness* between the average voice model and adapted target model was found in [10]. This result was specific to adult speech synthesis using gender-dependent initial models and the correlation was weak. Research by Yamagishi and Watts has led to the conclusion that HMM adaptation provides the best solution for child speech synthesis, due to its ability to perform well with limited data [17], [19]. However, there is a need to improve the methods for the selection of training speakers to train a gender-independent initial model such that HMM adaptation can be applied successfully in child speech synthesis and yield better quality synthetic child voices.

C. Objective measures of synthetic voice quality

Subjective listening tests are typically used to evaluate synthetic speech. Numerous objective methods to accurately predict subjective listening scores have been proposed over the last few decades. However, only a few have demonstrated the capability to do so [20]. No single objective method has proven to be sufficiently reliable to evaluate synthetic speech.

A popular objective measure of the accuracy of the spectral envelope of synthetic speech is the average mel-cepstral distance (MCD). The distance is computed as the Euclidean distance between the mel-cepstral parameters of two speech samples.

The objective measure used to quantify the accuracy of the F_0 contour generated by the model is the root-mean-square-error (RMSE) of $\log F_0$. Since the F_0 is only observed in voiced regions, the RMSE of $\log F_0$ is only calculated for these regions.

III. Experiments and results

This section describes the data, experimental set-up and the evaluation procedures that were followed during the investigation. MCD and RMSE of $\log F_0$ were used to determine how close each of the speakers and different speaker combinations are to the target speaker. The objective evaluations obtained in this manner were subsequently compared to the results of a formal listening test. The aim was to determine whether the results of the objective measures will correspond to the results of the subjective evaluation.

A. Speech data

The speech data used to train the average voice model in this study was selected from the CMU-ARCTIC database [21]. This database consists of six speakers, each of whom read the

978-1-5090-3334-5/16/$31.00 ©2016 IEEE

same sentence set of 1131 phonetically balanced sentences. The sentence set corresponds to approximately 1.5 hours of speech data per speaker.

Four speakers were selected from the CMU-ARCTIC database and used to train the average voice models for this study. The average models were adapted to a South African English male child target speaker using 100 sentences as adaptation data. The child data was collected during a previous study [9] and amounts to 8 minutes of speech.

B. Experimental setup

Four speaker dependent voices were built, one for each of the four training speakers. Each of these voices were then adapted to the target speaker. A US phoneset was used for both the training and target speakers. Using the four training speakers, each speaker was paired with every other speaker, resulting in six additional voices, of which four were gender-independent and two were gender-dependent. The speaker combinations and resulting voices are summarised in Table I.

TABLE I: Combinations of training speakers

Speaker 1	Speaker 2	Voice ID
bdl	rms	bdl_rms
bdl	clb	bdl_clb
bdl	slt	bdl_slt
rms	clb	rms_clb
rms	slt	rms_slt
clb	slt	clb_slt

C. Objective Evaluation

Forty test sentences were synthesized for each voice. The test sentences were taken from children stories that were not included in the training or adaptation data. The mel-cepstral distance and the RMSE for $\log F_0$ were calculated for each of the voices listed in Table I. The distance between the output generated by the initial model and the output generated by the adapted model was calculated. Dynamic Time Warping (DTW) was applied to ensure that the temporal differences between the two samples did not influence the distance measures.

1) Results: The objective distance measures corresponding to the speaker dependent and speaker independent models are shown in Tables II and III respectively.

TABLE II: Objective measures calculated for speaker dependent models

Speaker	Average MCD [db]	Average RMSE $\log F_0$ [cent]
bdl	2.94	55
rms	2.54	76
clb	2.12	14
slt	2.79	18

According to the results presented in Table II, the speaker who is closest to the target speaker in terms of mel-cepstral distance is clb followed by rms. In terms of RMSE of $\log F_0$, clb and slt were found to be the closest to the target speaker. The result obtained for RMSE for $\log F_0$ is as expected because children are known to have higher fundamental frequency ranges that are closer to those of female voices than male voices. Therefore, for the remainder of the study, only the MCD measures were used.

TABLE III: Objective measures calculated for speaker independent models

Voice	Average MCD [db]	Average RMSE $\log F_0$ [cent]
bdl_rms	2.11	68
bdl_clb	2.22	31
bdl_slt	2.36	36
rms_clb	1.79	45
rms_slt	2.01	47
clb_slt	2.05	16

Since [9] showed that the best initial model for child speech synthesis is a gender-independent model, the hypothesis tested in this study was that the closest female speaker and closest male speaker combination would result in the best gender-independent initial model. From the objective results presented in Table III, it is evident that the speaker combination that yields the smallest average MCD to the target speaker is rms_clb. This confirms that the hypothesis made is valid in terms of this objective measure.

In a similar manner it was expected that the clb_slt combination would result in the smallest RMSE F_0 distance and bdl_rms in the biggest distance between the average voice model and the target speaker. This hypothesis is also confirmed by the results in Table III. In both cases, the closest voices (rms_clb and clb_slt, respectively) are substantially different from the remaining voices.

D. Subjective Evaluation

Subjective evaluations were conducted using formal perceptual listening tests that were administered via a web interface. The participants were 19 adult listeners native to South Africa. The key properties that were evaluated included user preference, naturalness and intelligibility. The listening test therefore consisted of the following sections:

1) Paired comparison test to evaluate user preference.
2) Mean-Opinion-Score (MOS) test to evaluate naturalness.
3) Transcription test to evaluate intelligibility.

One hundred test sentences from children stories were synthesized using each speaker independent voice adapted to the child target speaker. The test sentences were taken from children stories that were not included in the training or adaptation data.

A paired comparison test was conducted to determine user preference. In this test, each participant listened to a total of 30 voice samples. Every question comprised of two samples, one from each of the six speaker combinations. In this way every voice combination could be compared with another voice at least twice in the total set of questions. All samples were ordered randomly. The listener was required to select either 'Sample A', 'Sample B' or 'Sounds the same'.

A MOS listening test was conducted to evaluate naturalness. The participants were required to listen to 33 voice samples in total of which five corresponded to one of the six speaker

978-1-5090-3334-5/16/$31.00 ©2016 IEEE

combinations. They were asked to rate each sample based on its naturalness. The rating was performed in terms of a 5-point scale. The points of the scale were defined as follows: 1 - Completely Unnatural, 2 - Unnatural, 3 - Slightly natural, 4 - Natural and 5 - Completely Natural. Among the samples, three natural samples of the target child speaker were included as a benchmark.

Intelligibility was evaluated by asking the participants to listen to the adapted speech samples and transcribe the audio. 30 semantically predictable sentences (SPS) were transcribed per listener, of which five random samples corresponded to each speaker combination model. Research on child speech synthesis has shown that children's speech is very difficult to interpret with semantically unpredictable sentences, which is the conventional form of sentences used in these types of tests. Therefore semantically predictable sentences were used during the evaluations conducted in [8]. Similarly, semantically predictable sentences were also used in this study. Using these transcribed sentences, average word error rate (WER) was calculated for each model. The transcription WER is calculated using the formula provided in the Blizzard 2007 challenge guidelines [22]. Spelling mistakes and typographical errors were corrected before the WER was calculated.

1) Results:

a) User preference: The results of the user preference test are summarised in Figure 2. The results indicate that rms_slt was the most preferred voice with many more votes than any of the other voices. The least preferred voice was bdl_clb.

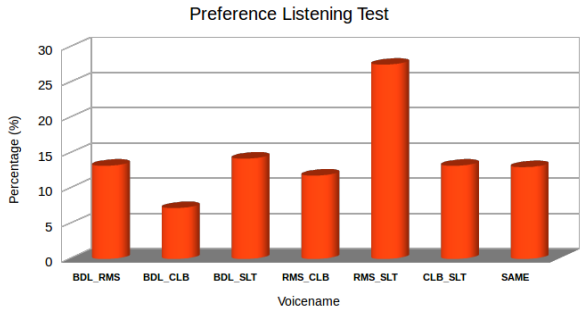

Fig. 2: Subjective evaluation: Preference listening test

b) MOS test: The results of the MOS test are presented in Figure 3 in a standard boxplot. The median is represented by a solid bar across the middle of the box, whiskers extend to 1.5 times the inter-quartile range and outliers are represented with circles.

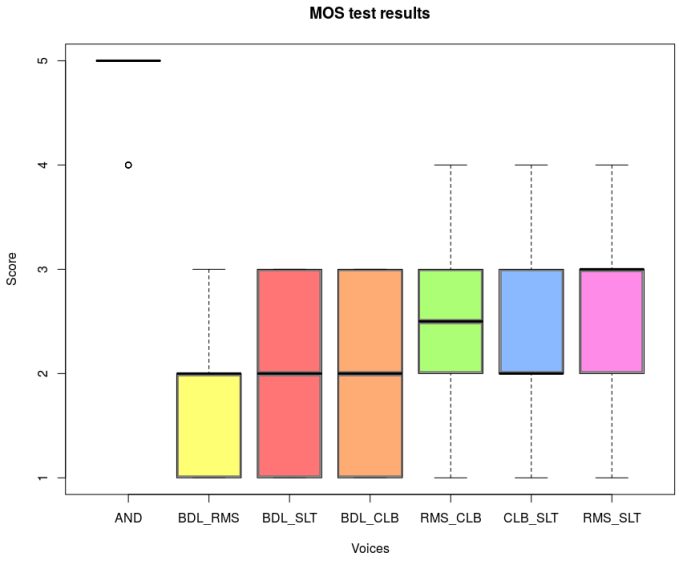

Fig. 3: Subjective evaluation: MOS test results

The results of the MOS test clearly indicate that the voice that was considered to be the most natural of the six voices is rms_slt with a median of 3 (which means the voice overall is considered to be slightly natural). Given that a few listeners evaluated the original recording of the target speaker as *natural* (which is a score of 4), the overall result of the adapted synthetic voice could be regarded as a positive one. The voice that performed the worst is bdl_rms. Even though it had a median of 2 like the other voices, majority of its votes lie in the region of 1 and 2.

c) Intelligibility: The results of the intelligibility test were quantified in terms of the word error rate (WER) that occurred in every transcribed sentence (with typographical errors and spelling errors corrected). The WERs corresponding to the different voices are illustrated in Figure 4. The results show that rms_clb performed the best, followed by rms_slt and clb_slt with a negligible difference between them. It is clear that rms_clb is the most intelligible as it leads with a 5% WER from the other two voices that follow. Bdl_rms performs the worst.

IV. Correlation between objective and subjective measures

An analysis of the two sets of results reveals that the objective results directly correlate with the results obtained for the intelligibility test. The rms_clb voice performed the best in the intelligibility test with a WER of 14.11%. It is the voice combination of the two closest speakers in the test set and individually they are also the closest male and closest female voices from the test set.

Rms_slt performed better than rms_clb in terms of naturalness but rms_clb follows closely. Rms_slt was also the most preferred voice but by closely analysing the results it was noted that the least preferred is voice is bdl_clb. Both these voices correlate directly with the results obtained in the naturalness

978-1-5090-3334-5/16/$31.00 ©2016 IEEE

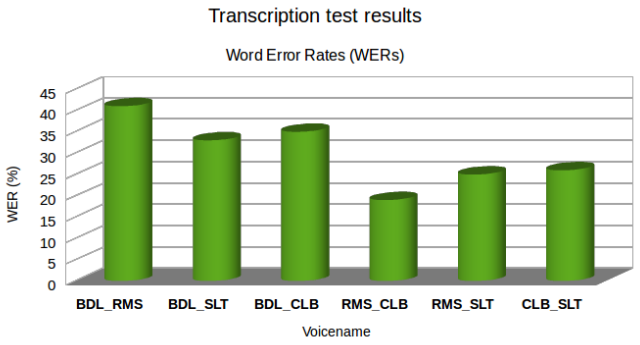

Fig. 4: Subjective evaluation: Transcription test results

test and thus it can be assumed that the preference test was biased towards naturalness.

Overall, when it comes to integrating these voices in real-world applications, the intelligibility of the voice will always be more important than the naturalness. In this case the voice that performed the best objectively outperformed the other voices in the subjective test for intelligibility and only fell slightly short in terms of naturalness.

V. Conclusion and Future Work

The results of this study revealed a correlation between the average mel-cepstral distance between the average voice model and the adapted voices and the overall intelligibility of the voices. Average MCD was also found to be somewhat correlated with the naturalness. This finding will be useful especially when attempting to synthesise child voices with a larger database of adult speech data. It will also improve the development of child voices because researchers can use this objective measure to determine whether specific training speakers can be clustered together to obtain average voice models that will improve the quality of the resulting child voice.

Further investigations are being conducted to determine whether the same measures can be used to select suitable training data from automatic speech recognition corpora [23] and whether the observed results generalise to other corpora of adult speech.

References

[1] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Acoustics, Speech and Signal Processing, 1996. ICASSP'96. IEEE International Conference on*, vol. 1, pp. 373–376, IEEE, 1996.

[2] J. Yamagishi, T. Kobayashi, M. Tachibana, K. Ogata, and Y. Nakano, "Model adaptation approach to speech synthesis with diverse voices and styles," in *Acoustics, Speech and Signal Processing, 2007. ICASSP'07. IEEE International Conference on*, vol. 4, pp. IV–1233, IEEE, 2007.

[3] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[4] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *EUROSPEECH*, vol. 5, pp. 2374–2350, September 1999.

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

[6] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, Y. Guan, R. Hu, K. Oura, Y.-J. Wu, *et al.*, "Thousands of voices for HMM-based speech synthesis – Analysis and application of TTS systems built on various ASR corpora," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 5, pp. 984–1004, 2010.

[7] J. Yamagishi, *Average-voice-based speech synthesis*. PhD thesis, Tokyo Institute of Technology, 2006.

[8] O. Watts, J. Yamagishi, K. Berkling, and S. King, "HMM-based synthesis of child speech," in *Proceedings of The 1^{st} Workshop on Child Computer and Interaction.*, (Crete, Greece), October 2008.

[9] A. Govender, F. de Wet, and J. R. Tapamo, "HMM adaptation for child speech synthesis," in *INTERSPEECH 2015*, (Dresden,Germany), pp. 1640–1644, September 2015.

[10] J. Yamagishi, O. Watts, S. King, and B. Usabaev, "Roles of the average voice in speaker-adaptive HMM-based speech synthesis," in *INTERSPEECH*, pp. 418–421, September 2010.

[11] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proc. of 6th ISCA Workshop Speech Synthesis*, pp. 294–299, August 2007.

[12] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[13] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Acoustics, Speech and Signal Processing, 2000. ICASSP'00. IEEE International Conference on*, vol. 3, pp. 1315–1318, IEEE, 2000.

[14] J. Yamagishi, H. Zen, Y.-J. Wu, T. Toda, and K. Tokuda, "The HTS-2008 system: Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge," in *Proc. Blizzard Challenge 2008*, September 2008.

[15] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP'92, IEEE International Conference on*, vol. 1, pp. 137–140, IEEE, 1992.

[16] D. Arifianto, T. Tanaka, T. Masuko, and T. Kobayashi, "Robust estimation of speech signal using harmonicity measure based on instantaneous frequency," *IEICE TRANSACTIONS on Information and Systems*, vol. 87, no. 12, pp. 2812–2820, 2004.

[17] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 1, pp. 66–83, 2009.

[18] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *Acoustics, Speech and Signal Processing, 1983. ICASSP'83. IEEE International Conference on*, vol. 8, pp. 93–96, IEEE, 1983.

[19] O. Watts, J. Yamagishi, S. King, and K. Berkling, "Synthesis of child speech with HMM adaptation and voice conversion," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 5, pp. 1005–1016, 2010.

[20] R. F. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Communications, Computers and Signal Processing, 1993., IEEE Pacific Rim Conference on*, vol. 1, pp. 125–128, IEEE, 1993.

[21] J. Kominek and A. W. Black, "The CMU arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[22] R. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the Blizzard challenge 2007 listening test results," *Proc. BLZ3-2007 (in Proc. SSW6)*, 2007.

[23] A. Govender, B. Nouhou, and F. de Wet, "HMM adaptation for child speech synthesis using ASR data," in *Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech), 2015*, pp. 178–183, IEEE, 2015.

978-1-5090-3334-5/16/$31.00 ©2016 IEEE