INTERSPEECH 2009 BRIGHTON

HMM adaptation and voice conversion for the synthesis of child speech: a comparison

Oliver Watts[1], *Junichi Yamagishi*[1], *Simon King*[1], *Kay Berkling*[2]

[1]Centre for Speech Technology Research, University of Edinburgh, UK
[2]Inline Internet Online Dienste GmbH, Germany

O.S.Watts@sms.ed.ac.uk jyamagis@inf.ed.ac.uk Simon.King@ed.ac.uk kay@berkling.com

Abstract

This study compares two different methodologies for producing data-driven synthesis of child speech from existing systems that have been trained on the speech of adults. On one hand, an existing statistical parametric synthesiser is transformed using model adaptation techniques, informed by linguistic and prosodic knowledge, to the speaker characteristics of a child speaker. This is compared with the application of voice conversion techniques to convert the output of an existing waveform concatenation synthesiser with no explicit linguistic or prosodic knowledge. In a subjective evaluation of the similarity of synthetic speech to natural speech from the target speaker, the HMM-based systems evaluated are generally preferred, although this is at least in part due to the higher dimensional acoustic features supported by these techniques.
Index Terms: child speech, statistical parametric speech synthesis, HMM-based speech synthesis, voice conversion, HTS, Average Voice Models, Festival

1. Introduction

The synthesis of child speech presents particular difficulties for data-driven systems due to the type and quantity of data which it is feasible to collect from child speakers. Data-driven speech synthesisers are conventionally trained on corpora that are phonetically balanced, consistently read, and cleanly recorded. The type of child speech typically available, in contrast, more closely resembles 'found' data. It will be unbalanced in terms of the units it covers due to the fact that it is infeasible to have a child read a recording script that has been specially designed with a view to phonetic/prosodic coverage. Children's speech shows a greater degree of variability than adults', and the data will typically be read less consistently than adult data. Finally, the data available will typically be imperfectly recorded, as practical considerations mean it is less easy to get a child into a purpose-built recording booth than an adult speaker.

In [1], what is to our knowledge the first data-driven synthesis of child speech was presented. In this work, the difficulties inherent in a small corpus of child speech were overcome by the use of Hidden Markov Model (HMM)-based synthesis techniques. In particular, speaker adaptation techniques were used to adapt an existing average voice model – trained on a clean, phonetically rich corpus collected from six adult speakers – to the speech characteristics of a child target speaker. We considered the resulting synthetic speech successful in that it clearly reflected the nature of the training data: the speech synthesiser sounds like a child reading, with the same patterns of hesitancy and disfluency observed in the training data.

However, there exist methods for imposing the voice characteristics of a novel speaker on an existing synthesiser besides those employed in the above work. Voice conversion techniques aim to convert the speech of a source speaker in such a way that the converted speech appears to have been produced by a different, target speaker. As well as being applied to the conversion of natural speech, popular voice conversion techniques such as those based on spectral conversion with Gaussian mixture models (GMMs: [2],[3]) have been used to convert speech synthesisers' output to the voice characteristics of new speakers. For example, in [4], the output of a concatenative (diphone) synthesiser is used to create 'source speaker' training data for a voice conversion model. The voice conversion model allows arbitrary novel utterances subsequently produced by the synthesiser to be converted to the voice characteristics of the target speaker.

The application of voice conversion to the output of an existing speech synthesiser is particularly attractive in the context of sample-based concatenative methods where there is no statistical model whose parameters can be transformed as in the case of HMM-based synthesisers. Additionally, voice conversion can be performed when annotation of the training data is limited: no knowledge of the linguistic/prosodic structure of the data is necessary as e.g. the classes of spectral features represented by the Gaussian mixture components are determined purely on the basis of acoustic measures.

Lack of dependence on knowledge of the data's linguistic and prosodic structure can be beneficial in some circumstances, and when the match between source and target speaker is close, acceptable results can be obtained by ignoring the conversion of duration altogether. However, the distinctive patterns of hesitancy and disfluency observed in the segmental durations of our child target speaker are a very distinctive part of the speaker's identity and are quite different from the patterns found in any adult speech database. It does not seem reasonable to expect to capture the speaker characteristics of this data in a satisfactory manner when performing speaker transformation while ignoring the conversion of duration. But the satisfactory conversion of speaker durational characteristics is not straightforward in a voice conversion setting where no use is made of linguistic and prosodic annotation.

The purpose of the present study is to compare the potential offered by HMM-based adaptation and GMM-based voice conversion techniques for the transformation of existing synthesisers trained on adult speech to the voice characteristics of our child target speaker.

Table 1: *Identifying letters used for each system. Transformations are to the target speaker in all cases except for the duration adaptation of system B. HMM: HMM adaptation (CSMAPLR + MAP), VC: Voice Conversion (including uniform stretch for duration).*

System identifier	Base system	Transformation method Spect.	F0	Dur.
A	HTS average voice	HMM	HMM	HMM
B	HTS average voice	HMM	HMM	HMM (to SLT)
C	Multisyn SLT	VC	VC	*None*
D	Multisyn SLT	VC	VC	VC

2. The systems built

2.1. Overview

The rationale of the investigation presented here was to compare the two types of system in as practical a way as possible. This meant constructing and comparing systems that would be credible in a real-world context. An adaptive HMM-based synthesiser (System A in Table 1) was matched against a voice-converted synthesiser (System C) which was based on an existing unit selection synthesiser. By doing this we in one sense reduced the usefulness of the evaluation in that base synthesiser type (unit selection / statistical parametric) was not a factor kept constant. In another sense, however, the evaluation was more useful than if System C had been based on an HMM-based system in that the resulting systems are more representative of the kinds of speech synthesiser that are actually being built. Accounts of voice conversion applied to concatenative systems are more easily found in research literature (e.g. [4]) than accounts of voice conversion applied to statistical parametric synthesisers.

Systems B and D represent a concession to the need to control for some of the differences between systems in the evaluation. As mentioned above, in GMM-based voice conversion, the source speaker's durations are typically used unconverted in the output speech and this is the case in our System C. In HMM-based adaptive systems, on the other hand (e.g. our system A), models are routinely adapted to speaker-specific duration characteristics. System D represents an attempt to bridge this gap by converting duration in a shallow, data-driven fashion. In addition to voice conversion techniques being applied to spectral features and F0, knowledge of the total duration of training data for source and target speakers is used to uniformly stretch converted utterances' duration by an appropriate factor. In System B, the opposite approach to lessening the same difference between systems A and C is followed. That is, the duration model used is not adapted to the target speaker, but to one of the adult speakers on whose data the average voice was trained (SLT, on whose data the unit selection voices underlying Systems C and D were also trained). The idea was to produce a system more easily comparable with System C where SLT's durations were also used in the converted speech.

2.2. Target speaker data collection

Collection and preparation of this data is described more fully in [1]. Briefly, the North American-accented English speech of a 7-year old tri-lingual (Spanish, English, German) female was collected using a headset microphone in an informal setting at the home of one of the authors over the course of several months. The subject was very familiar with the story book text, which she was allowed to read without interruption. A total of just over 100 minutes of speech data were collected.

The data was segmented into utterance-sized units and hand-transcribed with considerable care. A number of sentences were held out from the corpus for later use in evaluation.

The necessary linguistic and prosodic annotation was obtained using the Multisyn voice-building tools [5]. The forced alignment used to determine the final phonetic transcription allows for vowel reduction and the insertion of pauses between words, pause insertion being particularly important in the case of such hesitantly read data.

In the present experiments, only one size of target speaker dataset was used, consisting of the whole of the training corpus. We note that no very small dataset was used in the evaluation, and that it is on just such a set that we might expect voice conversion methods to outperform HMM-based ones.

2.3. HTS systems (Systems A and B)

System A was previously described in [1] (where it received the identifying letter F). System B was newly constructed for these experiments but the same procedure was used: it is merely a combination of elements taken from two different adaptive voices adapted in the same way but to different speakers (our child target speaker and the speaker SLT from the ARCTIC database). These systems were built using HTS version 2.1 [6].

2.3.1. Training

Both systems adopted the gender-mixed average voice from the HTS entry in the Blizzard Challenge 2007 [7]. Details of the training of this gender-mixed average voice model are given in [7]. Briefly, it was trained on the six adult speakers of CMU-ARCTIC speech database [8], four male and two female. First, the speech of the training corpus was parameterised as 40 mel-cepstral coefficients, log F0 and the energy of aperiodic components in 5 frequency bands, and the dynamic and acceleration features derived from all of these, yielding a 138-dimension observation vector for the HMMs. F0 was extracted using a three-stage procedure, where the outputs of three different pitch-trackers were made to 'vote' on the F0 value of any given frame. Spectral analysis was performed with the high quality vocoder STRAIGHT [9], and the STRAIGHT spectra were converted to mel-cepstral coefficients. After the data had been parameterised, two gender-dependent average voice models were trained using Speaker Adaptive training (SAT); that is, speaker normalisation was applied during estimation of the models, to avoid different speaker-dependent voice characteristics "diluting" the average models. Then, the parameters of both gender-dependent models were clustered and tied using decision-tree based clustering, with gender included as a context feature. Then the clustered HMMs were re-estimated using SAT, regression classes for the normalisation being determined from the gender-mixed decision-trees. State durations obtained during this estimation were used to initialise duration probability distributions which were then clustered. SAT was performed on the complete HSMMs to re-estimate all parameters (including duration) with speaker normalisation.

It should be noted that the device that enables the incorporation of high-level prosodic features into the model is the decision tree for context clustering. Context clustering enables the extreme context-dependency of speech units, which are defined not only in terms of immediate phonetic and linguistic contexts, but also in terms of long-range factors that might have a bearing on units' prosodic realisation (see [10] for the full list

used). The rich context-dependency of the speech units results in a very large number of possible models. This in turn means that almost all models will be sparsely represented in the training corpus and that, at synthesis time, models of missing units will certainly need to be created. Both of these problems are solved by mapping the large number of possible HMM states onto a much smaller number of states whose parameters are actually estimated. The mapping from logical to physical HMM states is done by means of decision-trees.

2.3.2. Adaptation

Adaptation to the target speakers was performed with the procedure used for the same HTS entry in the Blizzard Challenge. Adaptation was performed with a combination of constrained structural maximum a posteriori linear regression (CSMAPLR), using the decision tree obtained during average-voice training to structure the hierarchical transforms, followed by maximum a posteriori (MAP) adaptation.

System A is made up of the spectral, F0, aperiodicity and duration models adapted in this way to the child target speaker. The only difference in System B is that the duration model has been substituted for one adapted to the data of the adult speaker SLT.

2.3.3. Synthesis

Ten sentences from the story *Goldilocks and the Three Bears* which were to be used in evaluation were synthesised with the Festival speech synthesis system [11]. Festival's front-end performed the phonetic and linguistic predictions needed to provided a sequence of context-dependent labels for each utterance. Based on these predictions, parameters were generated using the two systems that had been trained, and waveforms were synthesised from those parameters.

2.4. Voice conversion systems (Systems C and D)

2.4.1. Training

The synthesiser used as the 'source speaker' in the voice conversion systems (Systems C and D) was an existing unit selection voice which had been built with the data of the speaker SLT from the ARCTIC database. The voice had been constructed using the Multisyn voice building tools ([5]).

The voice conversion systems were built using scripts available as part of the FestVox scripts implementing techniques developed by Toda ([3], [12]). First of all, the missing half of the necessary parallel corpus of source and target speaker speech was created by synthesising the contents of the child speech corpus using the base unit selection synthesiser. Then the speech was parameterised, using the same procedure described in Section 2.3.1 above with STRAIGHT analysis rather than the FestVox analysis tools. The only departure from the analysis procedure described in Section 2.3.1 was that a lower dimension static feature vector was used (24 – the 0th coefficient was not used for training GMMs). This was a result of initial work in which attempts at training joint GMMs on much higher order features failed even with few mixture components.

The conversion model for spectral features was trained as follows. The static parameters were supplemented by dynamic features and then joint feature vectors were obtained from source and target speech by alignment with Dynamic Time Warping. The parameters of a GMM (weights, means and covariance matrices for 128 mixture components) over the joint features were initialised using Vector Quantization, and then iteratively refined using Expectation Maximisation. The data alignment and GMM training were iterated.

The conversion model for F0 was obtaining by computing the mean and standard deviation of both source and target speakers' log F0. Additionally, the necessary information for performing rudimentary duration conversion by uniform stretching was collected in the form of a duration scaling factor.

2.4.2. Synthesis and conversion

The sentences to be used in evaluation were synthesised with Festival's front-end as in Section 2.3.3, but this time waveform generation was performed by the concatenation of units selected from the SLT database by the Multisyn unit selection engine. The resulting waveform was then analysed in the same way as the training corpus had been. The spectral features were supplemented with dynamic features, and a sequence of single mixture component conditional probability density functions was determined from the GMM and input speech vectors using Viterbi selection. These PDFs were used to compute maximum likelihood static parameter sequences considering both static and dynamic parts of the distributions.

Source speaker's log F0 was converted by normalising the log F0 contour using the source speaker's mean and standard deviation, and then imposing the target speaker's mean and standard deviation (computed during training) on the resulting contour.

Speech was then resynthesised using the source speaker's power and aperiodicity measures unmodified together with the converted spectral features and log F0.

For system D, the additional step of converting duration was performed by uniformly stretching the converted utterance in accordance with the duration scaling factor computed during training. Utterances' duration was scaled using Pitch Synchronous Overlap and Add (in Praat: [13]).

3. Subjective Evaluation

3.1. Evaluation procedure

An XAB test was conducted in which a pairwise comparison was made of the four systems in terms of the similarity of the synthetic speech to the natural speech of the target speaker. Four reference sentences spoken by the target speaker which had been held out of the training corpus were analysed and resynthesised as described in Section 2.3.1 above with no manipulation of the features. They were presented at the beginning of the evaluation and at intervals throughout it as X, and listeners were encouraged to listen to the samples as much as they wanted. The ten 'Goldilocks' sentences were synthesised with each of the four systems, and for each sentence an AB pair (randomly ordered) was made for each pair of systems, resulting in 60 AB pairs. The listening test was conducted via a web browser, with a total of 10 unpaid listeners. The 60 pairs were presented in random order and listeners were asked to choose the sentence in which the synthetic speech's speaker characteristics were most similar to those of the natural reference samples.

3.2. Evaluation results

Figure 1 shows the results of the evaluation. Significant preferences were detected for all pairs of systems except B vs. D. System A – the HMM synthesiser with duration adapted to the child target speaker – was preferred to both voice conversion

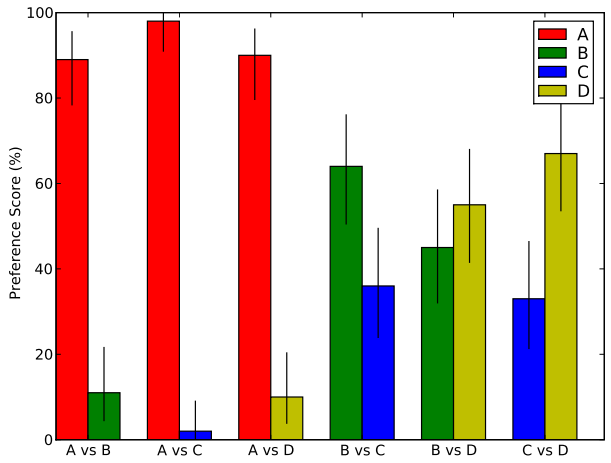

Figure 1: *Results of XAB test for speaker individuality. Vertical lines show 95% confidence intervals (with Bonferroni correction).*

systems and to the other HMM-based synthesiser.

The evaluation shows that the HMM-based systems were generally preferred as more similar to the original speaker than the voice converted unit selection systems. In only one case (B vs. D, where the HMM-based system not adapted to target speaker duration was compared with the voice-converted system with uniform duration modification) was no significant preference for the HMM-based system found.

4. Conclusions

The premise of the experiment presented here was that the two types of system under investigation should be compared in a 'realistic' a way as possible. The considerable differences between the systems were not factored out in the construction of stimuli because these differences were motivated by real-world considerations. That is, a concatenative system was used as the basis of the voice conversion systems (rather than a statistical parametric synthesiser which would have made comparison with the other systems easier) because the possibility of performing voice conversion is more relevant in the context of concatenative systems. Likewise, higher order acoustic features were used for the HMM-based systems because the method supported higher order features; lower order features were used for performing voice conversion due to the fact that attempts at using higher order features in initial work resulted in difficulties training the GMMs. Both discrepancies are reasonable according to the rationale with which the experiment was designed in that two different but credibly real-world systems were compared using the highest order acoustic features they were found to support in initial tests. The evaluation revealed a general preference for the HMM-based systems. However, the discrepancies in the construction of the different types of system mean that it is not possible to attribute the superior performance of the HMM-based systems used directly to the use of HMM-based adaptation. The superior quality of waveform synthesis from higher order acoustic features undoubtedly contributes something to the relative success of the HMM-based systems. We hypothesise that if differences of synthesiser type (concatenative / statistical parametric) and acoustic feature order were controlled for in evaluation, synthetic speech from an adapted HMM-based synthesiser would still outperform speech from a voice-converted system due to the high-level linguistic and prosodic information which can be used to inform the adaptation. We plan to test this hypothesis in a future evaluation.

In ongoing work, we are exploring the effect of reducing the amount of target speaker training data on the relative performance of HMM and voice conversion based methods. It is hypothesised that listener preference for HMM-based systems over voice conversion systems will be reduced when less target speaker data is available.

5. Acknowledgements

This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF). (http://www.ecdf.ed.ac.uk/). The ECDF is partially supported by the eDIKT initiative (http://www.edikt.org.uk). The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project `http://www.emime.org`). JY is partially supported by EPSRC. SK holds an EPSRC Advanced Research Fellowship. We thank Robyn Zundel, daughter of Kay Berkling, for her contribution to this research through many hours of reading and recording. We also thank the anonymous reviewers for some helpful comments.

6. References

[1] O. Watts, J. Yamagishi, K. Berkling, and S. King, "HMM-based synthesis of child speech," in *Proc. of The 1st Workshop on Child, Computer and Interaction (ICMI'08 post-conference workshop)*, Crete, Greece, Oct. 2008.

[2] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *Speech and Audio Processing, IEEE Transactions on*, vol. 6, no. 2, pp. 131–142, Mar 1998.

[3] T. Toda, A. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2222–2235, Nov. 2007.

[4] A. R. Toth and A. W. Black, "Incorporating durational modification in voice transformation," in *Interspeech 2008*, Sep. 2008.

[5] R. A. J. Clark, K. Richmond, and S. King, "Multisyn: Open-domain unit selection for the Festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317–330, 2007.

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

[7] J. Yamagishi, H. Zen, T. Toda, and K. Tokuda, "Speaker-independent HMM-based speech synthesis system — HTS-2007 system for the Blizzard Challenge 2007," in *Proc. BLZ3-2007 (in Proc. SSW6)*, Aug. 2007.

[8] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Proc. ISCA SSW5*, 2004.

[9] H. Kawahara, I. Masuda-Katsuse, and A. 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.

[10] H. Zen, K. Tokuda, and T. Kitamura, "An introduction of trajectory model into HMM-based speech synthesis," in *Proc. ISCA SSW5*, 2004.

[11] A. Black, P. Taylor, and R. Caley, *The Festival Speech Synthesis System*, University of Edinburgh, 1999.

[12] A. W. Black and K. A. Lenzo, "Building synthetic voices," 2007. [Online]. Available: http://festvox.org/bsv/

[13] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2005. [Online]. Available: http://www.praat.org/