CREATING SYNTHETIC VOICES FOR CHILDREN BY ADAPTING ADULT AVERAGE VOICE USING STACKED TRANSFORMATIONS AND VTLN

Reima Karhila[†]*, D. R. Sanand*[‡]*, Mikko Kurimo*[†] *and Peter Smit*[†]

[†] Adaptive Informatics Research Center, Aalto University, Finland.
[‡] Department of Electronics and Telecommunications, NTNU, Trondheim, Norway.
reima.karhila@aalto.fi,rama.doddipatla@iet.ntnu.no,mikko.kurimo@aalto.fi,peter.smit@aalto.fi

ABSTRACT

This paper describes experiments in creating personalised children's voices for HMM-based synthesis by adapting either an adult or child average voice. The adult average voice is trained from a large adult speech database, whereas the child average voice is trained using a small database of children's speech. Here we present the idea to use stacked transformations for creating synthetic child voices, where the child average voice is first created from the adult average voice through speaker adaptation using all the pooled speech data from multiple children and then adding child specific speaker adaptation on top of it. VTLN is applied to speech synthesis to see whether it helps the speaker adaptation when only a small amount of adaptation data is available. The listening test results show that the stacked transformations significantly improve speaker adaptation for small amounts of data, but the additional benefit provided by VTLN is not yet clear.

Index Terms— Speech synthesis, Adaptation, Child Speech, VTLN and Stacked Transformations.

1. INTRODUCTION

Speaker-adaptive Hidden Markov Model (HMM)-based speech synthesis, where an average voice model is adapted to a new speaker with very little training data is going to have a large impact on speech technology applications. Users can have systems speaking in their own voice or in the person's voice of their choice. A typical HMM based speech synthesis system consists of statistical models to represent the acoustic and prosodic characteristics that are used for synthesising speech in a source-filter fashion[1]. Adaptive synthesis is based on models trained from a large population of speakers and that form a representative average voice model for this entire population. This average voice model is adapted to a particular speaker of interest using a few utterances spoken by that speaker, which is later used for synthesising speech for that speaker.

Creating a good single-speaker voice model requires at least a few hours of training data from that speaker, whereas adapting the average voice to a particular speaker requires around 3 minutes of recorded speech for generating good quality voice [2]. It is interesting to note that as few as five sentences are sufficient to make a voice resemble the target speaker and is enough for listeners to recognise the correspondence between the natural and synthetic voices[3].

It is common to use an average voice model that is built from a representative sample of speakers in a population (for example: "Finnish adults") to synthesise speech for a speaker having similar voice characteristics of the population. A recent study discusses the relation between the average voice model and the target voice for adaptation. It has been shown that, the more the voices differ, the worse the quality of the resulting adapted voice [4]. There have been very few studies to understand how an average voice model adapts to speakers in a different population group. Such a situation arises when we do not have sufficient training data to create a robust average voice model, but do have enough training data for speaker adaptation. An example for such a situation will be to adapt the adult average voice model to synthesise a child voice, or to synthesise a heavily accented speaker from an unaccented average voice model.

This paper focuses on synthesising a child's voice given an average voice model trained on the adult data. Training a synthesiser for children's speech is known to be challenging, especially because of the difficulties in obtaining a satisfactory amount of phonetically balanced training data [5, 6]. Because of this, HMM-based speech synthesis is well suited for the task, as the missing phonetic models can be estimated from available data quite reliably.

In this paper, we present investigations on generating child voices by adapting average voice models. We will show that adapting a high-quality average adult voice model or an average voice model built using a limited amount of child training data may not be ideal for synthesising child voices. In order to improve the synthetic quality of child voices using an adult average voice model, we propose an approach called stacked transformations. The idea is to create a cascade (or stack) of transforms, where we first adapt the average adult voice model to a representative child average voice model which is further adapted to a given target child speaker of our interest. We will show that the proposed approach provides very good synthetic voices for children when compared to directly adapting the average voice model. We also investigate the use of vocal tract length normalisation (VTLN) with stacked transformation to assess whether it can provide any improvement when there is very little adaptation data available from the target speaker. We present listening experiments to support our claims.

The rest of the paper is organised as follows: first we present the idea of stacked transforms. We then present the idea of using VTLN in speech synthesis. Later, we present the setup used in our experiments followed by description of the listening task. We then present our analysis followed by our conclusion and future work.

2. STACKED TRANSFORMATIONS

The idea of stacked transforms is to use multiple transforms in speaker adaptation, where each transform has a different role to play. The idea was presented in [7], where a cascade (or stack) of

This work was done while D. R. Sanand was at Aalto University.

The research leading to these results was partly funded from the Tekes project 40464/10 and Academy of Finland project 129674.

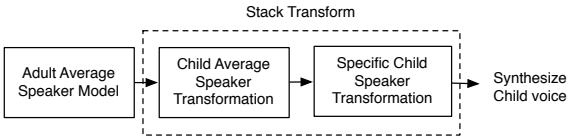

Fig. 1. Illustrating the idea of stacked transformation for synthesising child voice give the adult average speaker model.

transforms was used to perform speaker adaptation to handle foreign accented speech in automatic speech recognition (ASR). In its original form, the idea was to adapt a speaker-independent average voice model to an accented target speaker in several steps. First, the average voice model was adapted to an accent-specific average model estimated from a group of speakers sharing this particular accent. Then a traditional speaker adaptation transform was applied to further adapt the model set to the target speaker.

We follow a similar idea in our synthesis system, but instead of accented speakers, we use a group of child speakers to help synthesise the voice for a single target child speaker. The high-quality adult speaker-independent average voice is first adapted to an average child voice model. The resulting synthetic voice sounds child-like in terms of pitch, pronuncation and rhythm. This average child model is then further adapted to a specific target child speaker. The block diagram shown in Fig. 1 illustrates the idea of stacked transform in synthesis of child voice from the adult average voice model.

The obvious advantage of stacked transformations in speech synthesis lies in the use of a large corpus to train an initial voice model. This model has a good coverage of phonetic contexts and robust model clustering and selection. With a transform trained from a group of speakers, the model provides a solid basis for speaker adaptation. Other advantages of stacked transformations include the possibility of pre-training the domain transformations and reducing the need for storage space compared to a domain model by itself.

3. VTLN ADAPTATION FOR SPEECH SYNTHESIS

VTLN is widely used in ASR for normalising speaker variability that arises due to differences in vocal tract lengths of speakers uttering the same sound. The normalisation is achieved by scaling the spectra of speakers. VTLN is a simple approach and requires the estimation of a single parameter that controls the amount of spectral scaling. In practice, the scale- or warp-factor is estimated using a maximum likelihood based grid search over a pre-defined range of warp-factors and is given by:

$$\hat{\alpha_i} = \arg\max_{\alpha} \Pr(\mathbf{X}_i^{\alpha}|\mathbf{W}_i;\lambda) \tag{1}$$

where, $\mathbf{X}_i^{\alpha}$ represent the VTLN warped cepstral features, λ is the baseline model and $\mathbf{W}_i$ being the transcription of the data.

Unlike maximum likelihood linear regression (MLLR) based approaches that require sufficient data to robustly estimate all the elements of the transformation matrix, VTLN requires very little training data to optimally estimate a single warp-factor.

VTLN has been applied to HMM based speech synthesis [8] and has been shown to improve the synthetic speech quality when combined with adaptation based approaches [9, 10]. Using VTLN as a linear transformation eliminates the need to store warped features. The linear transform is derived by using a bilinear transform, where VTLN is considered as a filtering operation.

In this paper, we propose an alternative approach to derive the linear transformation for VTLN using the ideas of band-limited interpolation. This is based on the idea presented in [11], where the linear transformation was derived for MFCC features. We modify the derivation accordingly to suit Mel-generalised cepstral coefficient (MCEP) features. The MCEP feature extraction is summarised in Fig. 2. The mel-warping is done using an iterative technique as discussed in [12] using the plain cepstral values obtained after performing the inverse Fourier transform (IFFT).

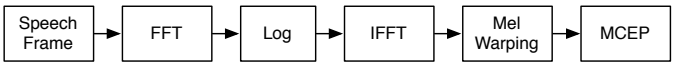

Fig. 2. Stages in MCEP feature extraction

The interpolation matrix to obtain VTLN warped MCEP in our approach is given by:

$$\mathbf{A}^{\alpha} = \frac{2}{N}\left[\mathbf{U}_{jk}^{\alpha}.\mathbf{V}_{ki}\right] \tag{2}$$

where N is the size of fast Fourier transform (FFT) and M represents the size of MCEP features. The matrices $\mathbf{U}^{\alpha}$ and $\mathbf{V}$ are given by:

$$\mathbf{V}_{ki} = \left[a_i\ cos(2\pi\beta_i k)\right]_{\substack{0 \le k \le M-1,\\ 0 \le i \le N-1}} \quad \beta_i = \frac{\nu_i}{2\nu_s}$$

$$\mathbf{U}_{jk}^{\alpha} = \left[a_j\ cos(2\pi\beta_j^{\alpha} k)\right]_{\substack{0 \le j \le N-1,\\ 0 \le k \le M-1}} \quad \beta_j^{\alpha} = \frac{\nu_j^{\alpha}}{2\nu_s}$$

and

$$a_i, a_j = \begin{cases} \frac{1}{2}, & i,j = 0, N-1 \\ 1, & i,j = 1,2,\ldots,N-2 \end{cases}$$

Here β_i represent the normalised frequency values after Mel-warping and β_j^{α} represent the Mel- and VTLN-warped normalised frequencies. ν_s represents the sampling frequency of the speech signal. ν_i and ν_j^{α} represent the Mel-warped and Mel- and VTLN-warped frequencies respectively. The problem is to reconstruct the values at β_j^{α} given the values at β_i and we do this using band-limited interpolation. The main advantage with this approach is that any frequency warping function can be used for VTLN scaling by choosing the appropriate values for β_j^{α}. It is important to realize that VTLN scaling is performed in the linear frequency Hertz (Hz) domain and not on the Mel-warped frequency domain. So when choosing the frequencies of β_j^{α}, the VTLN warped frequencies are calculated by first performing inverse Mel-warping followed by VTLN-warping and again Mel-warping.

4. EXPERIMENTS

4.1. Corpora

The Finnish SpeeCon corpora is used for training and testing the models. The corpora consists of speech data from both adults and children. The child data has 50 speakers with 60 utterances per speaker. The speakers are equally divided into age groups between 9-10 years and 11-12 years old[1] and having equal contributions from boys and girls. For the experiment, the speakers were divided into a training set of 40 speakers (total of 2367 sentences) and a test set of 10 speakers, selected randomly and as evenly as possible from the younger and older, and boy and girl groups. The adult average voice model used in the experiments was trained using 11057 utterances from 310 speakers, consisting of 147 female and 163 male speakers.

[1]The documentation specifies age groups 8-10 and 11-14, but older and younger speakers are not present in the corpus.

Table 1. Phonetic and contextual richness in training data. Start and end silences have been excluded.

Training set	Child	Adult
Nr. of speakers	40	310
Utterances	2 367	11 057
Unique quinphones	14 658	102 556
Unique contexts	28 946	449 815
Realized phones	84 533	706 608

The sentences spoken by children in SpeeCon had been collected from storybooks. Sentences had been shortened if necessary and difficult words had been removed. An overview of the phonetic coverage and complexity of the sentences is shown in Table 1. It is clear that the phonetic coverage of the child database is much weaker than that of the adult database, and this is likely to influence the synthesis performance of the trained model sets.

4.2. Model training

The adult and child average voice models were trained from the data described above using the same methods and tools as the EMIME 2010 Blizzard Entry[13]. In short, context-dependent multi-space distribution hidden semi-Markov Models (MSD-HSMMs) were trained on acoustic feature vectors comprising STRAIGHT-analysed Mel-generalised cepstral coefficients (MCEP), fundamental frequency and aperiodicity features. Speaker-adaptive training is applied to create a speaker-adaptive average voice model.

For stacked transformations, an average child speaker transformation was trained as a decision-tree based CSMAPLR transform [14] of all feature streams and duration. Two different stacked transforms sets were trained: A plain stack with no speaker adaptive training (St), and a VTLN stack, with the training data normalised with VTLN (StV). Speaker adaptation was done also as a CSMAPLR adaptation of all feature streams and duration. HTS tools were used to train the transforms and VTLN is also implemented in the same framework.

4.3. Listening Test

For the listening test, we chose three target speakers randomly: 10- and 11-year old girls and a 9-year old boy out of the test set of ten speakers. Table 2 summarises the listening experiment performed. We have four different types of average voice models to synthesize the child voice for the above mentioned speakers, i.e **Ad**, **Cd**, **St** and **StV**. Other than this, we also vary the amount of adaptation data used to adapt the average voice model for a specific speaker using 3, 10 and 50 utterances. This means, we have three sets of synthetic voices for each of the average speaker models and all together 12 synthetic voices for a specific speaker using all the average speaker models. The synthesis samples were generated from the child corpus's sentence prompts that were not used in training.

The listening test consisted of two tasks, and up to three repetitions of both tasks with different target speakers. 26 listeners started the listening test. Most listeners finished all three repetitions of both tasks, for a total of 67 data points for both tasks.

4.4. Task 1: Choosing the Best

In this task, the listeners were presented with a reference sample (natural speech) and two test samples that are synthesised using any of the average voice models specified above. The listener was asked

Table 2. Stimuli types in listening test

NS	Natural speech
VN	Vocoded natural speech
Ad	Average voice trained from **A**dult **d**ata
Cd	Average voice trained from **C**hild **d**ata
St (Stack)	**Ad** adapted using training data of **Cd**
StV	**Ad** adapted using training data of **Cd** after **V**TLN normalisation.
AdA$_n$	**Ad A**dapted with n sentences (3, 10 or 50)
CdA$_n$	**Cd A**dapted with n sentences (3, 10 or 50)
StA$_n$	**St A**dapted with n sentences (3, 10 or 50)
StVA$_n$	**StV A**dapted with n sentences (3, 10 or 50) after VTLN normalisation.

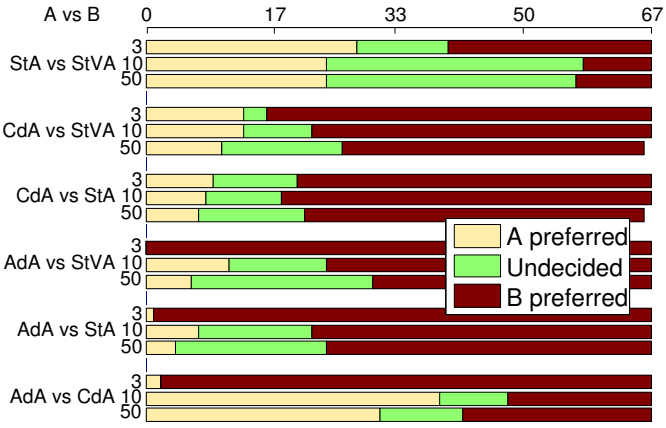

Fig. 3. Listening test results for task 1.

to answer: *"Which sample would you prefer to represent the reference?"*, with possible answers of sample A, sample B and *"I cannot tell the difference"*. The test samples were drawn from the available pool of four voices obtained using the different adaptation techniques (AdA, CdA, StA, StVA), but always use the same amount of adaptation data. Each sample in the pool was tested against every other sample, which results in 6 combination of tests for each voice for varying amount of adaptation data (3,10 and 50 utterances). So, we have a total of 18 tests for each test speaker and always using an identical sentence prompt for synthesis.

Figure 3 shows the results for the preference task. The synthetic voice generated using the stacked transformation systems (StA and StVA) was preferred by majority of the listeners when compared with the synthetic voice generated by directly adapting the adult average voice (AdA) or the child voice (CdA). This is statistically significant in all cases. It is also interesting to note that, the adaptation of an adult average voice seems to be preferred to adaptation of child voice when there is enough adaptation data. On the contrary, CdA is preferred over AdA when very little data is available.

The role of VTLN adaptation is hard to judge from these results. We do notice that the stacked system without VTLN (StA) is preferred over (StVA) when more adaptation data is available, although a lot of the listeners found it hard to make the judgement. In the 3 sentence case, the systems are equal. This is not yet enough to recommend using or not using VTLN, and an investigation into listener preference with even smaller amounts of data might be interesting, if not entirely useful.

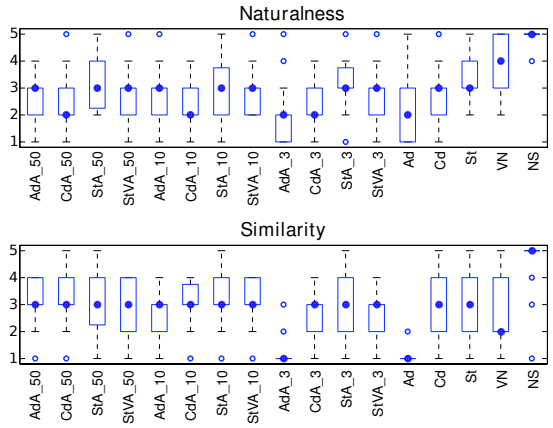

Fig. 4. Listening test results for task 2 in standard box plot format. Median marked as a solid circle, box extends to 25th and 75th percentiles and whiskers to the furthest data points that are not considered outliers. Outliers are marked with hollow circles.

4.5. Task 2: Naturalness and Similarity

In this task, the listeners were presented with a reference sample and all the synthesized voices from all the average speaker models for that speaker. The listener is asked to rate for each voice on a scale of 1-5 for similarity to the reference and the also for naturality. Each task was repeated for all the three test speakers. Listeners got the stimuli within each task in random order.

The results presented in Figure 4 show clearly the effects of varying amount of adaptation data. With 50 sentences, there is hardly any difference between the four investigated adaptation methods. When little data is available, it is not possible to create a satisfactory voice directly from the adult voice. Some of the task 1 results are further confirmed: Stacked transformations give better adapted voices than directly adapting the child average voice. An interesting detail is that the listeners find the average voices Cd and St very similar to the reference voices, something that is not typically seen in the case of adapting to adult voices. As there is typically more style variation in recordings of child speech, it is possible that the listeners are more willing to accept variation in synthetic child voices than when evaluating synthetic adult voices.

It should be noted, that vocoded samples of child speech seem to have more artifacts than vocoded adult speech. This raises questions about the choice of vocoders. STRAIGHT does not perform at its best with Finnish voices and it is possible that these voice properties are more prevailing in Finnish children's speech. Also, the synthesised sentences matched the training data of the children's average voice very closely. Test sentences synthesised from long and complex sentences revealed weaknesses in the performance of the child average voice, and this might be an object for another study.

5. CONCLUSIONS

Collecting enough data to build good-quality synthesis voices from children can be a daunting task. When there is limited amount of adaptation data, it is important that the average voice used as the basis for adaptation is of high quality and in the same domain as the target speaker. When there are not enough resources to collect speech data to build such a voice for children, it is possible to use speech data from adults as a basis and to create a high-quality average child voice via adaptation.

We have introduced stacked transformations and VTLN adaptation to the Finnish HMM-based speech synthesis framework and shown that with these tools we can create better speaker-adapted voices for children.

6. REFERENCES

[1] Heiga Zen, Keiichi Tokuda, and Alan W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039 – 1064, 2009.

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

[3] Reima Karhila and Mirjam Wester, "Rapid adaptation of foreign-accented HMM-based speech synthesis," in *Proc. Interspeech*, 2011.

[4] Junichi Yamagishi, Oliver Watts, Simon King, and Bela Usabaev, "Roles of the average voice in speaker-adaptive HMM-based speech synthesis," in *Proc.Interspeech*, 2010.

[5] Oliver Watts, Junichi Yamagishi, Kay Berkling, and Simon 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.

[6] 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, July 2010.

[7] Peter Smit, "Stacked transformations for foreign accented speech recognition," M.S. thesis, Aalto University School of Science, 2011.

[8] M. Hirohata, T. Masuko, and T. Kobayashi, "A study on average voice model training using vocal tract length normalization," Tech. Rep., IEICE, 2003.

[9] Lakshmi Saheer, Philip N. Garner, John Dines, and Hui Liang, "VTLN adaptation for statistical speech synthesis," in *Proceedings of ICASSP*, 2010.

[10] Lakshmi Saheer, John Dines, Philip N. Garner, and Hui Liang, "Implementation of VTLN for statistical speech synthesis," in *Proceedings of ISCA Speech Synthesis Workshop*, 2010.

[11] D. R. Sanand and S. Umesh, "Study of jacobian compensation using linear transformation of conventional MFCC for VTLN," in *Proc. Interspeech*, 2008.

[12] Keiichi Tokuda, Takao Kobayashi, and Satoshi Imai, "Recursive calculation of mel-cepstrum from LP coefficients," Tech. Rep., Nagoya Institute of Technology, 1994.

[13] Junichi Yamagishi and Oliver Watts, "The CSTR/EMIME HTS system for Blizzard Challenge," in *Proc. Blizzard Challenge,*, Kyoto, Japan, 2010.

[14] Yuji Nakano, Makoto Tachibana, Junichi Yamagishi, and Takao Kobayashi, "Constrained structural maximum a posteriori linear regression for average-voice-based speech synthesis," in *Proc. Interspeech*, 2006.