# HMM-based synthesis of child speech

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, October 2008.

# HMM-Based Synthesis of Child Speech

Oliver Watts
University of Edinburgh, UK
O.S.Watts@sms.ed.ac.uk

Junichi Yamagishi
University of Edinburgh, UK
jyamagis@inf.ed.ac.uk

Kay Berkling
Polytechnic Univ. of Puerto Rico
kay@berkling.com

Simon King
University of Edinburgh, UK
Simon.King@ed.ac.uk

## ABSTRACT

The synthesis of child speech presents challenges both in the collection of data and in the building of a synthesiser from that data. Because only limited data can be collected, and the domain of that data is constrained, it is difficult to obtain the type of phonetically-balanced corpus usually used in speech synthesis. As a consequence, building a synthesiser from this data is difficult. Concatenative synthesisers are not robust to corpora with many missing units (as is likely when the corpus content is not carefully designed), so we chose to build a statistical parametric synthesiser using the HMM-based system HTS. This technique has previously been shown to perform well for limited amounts of data, and for data collected under imperfect conditions. We compared 6 different configurations of the synthesiser, using both speaker-dependent and speaker-adaptive modelling techniques, and using varying amounts of data. The output from these systems was evaluated alongside natural and vocoded speech, in a Blizzard-style listening test.

## 1. INTRODUCTION

Child speech presents particular difficulties for data-driven speech synthesis due to the type and quantity of data which it is feasible to collect. Three characteristics are desirable in a corpus for constructing high quality synthetic voices: phonetic coverage, consistency, and quality of recording. Firstly, the material recorded should contain as wide a coverage of speech units in different phonetic and prosodic contexts as possible. This can be achieved by gathering a very large corpus of text and automatically extracting a sub-corpus in such a way that the phonetic coverage of the sub-corpus is optimised (e.g., [9]). Phonetically well-balanced recording scripts resulting from such methods are typically not coherent texts that children could be persuaded to read. It is more feasible to us, for example, story books familiar to the child, as is done in this work. This will typically result in a corpus with poor phonetic coverage.

Secondly, consistent speech quality is important in a speech synthesis database. The vocal and emotional control necessary to minimise inconsistency during and between recording sessions does not come readily to children. Use of a coherent 'script' enjoyable to the speaker increases the effects of emotional engagement with what is being read, which may be problematic from a speech synthesis perspective. In the present case, it led at times to fluctuations in speech quality and a variable reading style very different from that generally favoured in speech synthesis corpora.

Thirdly, the data must be well recorded, free from reverberation and background noise and with consistent acoustic quality. This is straightforward to achieve in the recording studio; however, it is more difficult to get a child into the studio than a paid voice talent. In the current work, the recordings were all made in the child's home and consequently contain considerable background noise, including page turns.

We addressed these problems, inherent in the recording of children's speech, by the use of HMM-based speech synthesis combined with particularly careful preparation of the data. HMM-based speech synthesis offers comparable quality to typical unit selection and concatenation systems [5]. However, it also offers two important capabilities that concatenative methods do not: an integrated data-driven method for dealing with missing units, and speaker adaptation. In the work reported here, we explore the potential of both of these capabilities for synthesising child speech. As far as we are aware, this is the first time HMM-based synthesis has been applied to child text-to-speech.

Speaker adaptation has become a key technique in many automatic speech recognition (ASR) systems. Methods from the Maximum Likelihood Linear Regression (MLLR) and Maximum A Posteriori (MAP) families are used to transform or adapt the parameters of already-trained HMMs, such that the likelihood of some *adaptation data* are increased. Speaker adaptation techniques have been used to adapt speech recognisers trained on adults' speech to the task of recognising children's speech, with some success. For example, part of the procedure described in [10] uses Maximum Likelihood-based model adaptation to adapt acoustic models trained on adults to child target speakers. In [3], Structural MAP Linear Regression is used for the same purpose. In speech synthesis, these methods have been used to adapt HMMs to new speakers using a very limited amount of data [15] – far less data than would be required to build a concatenative system, for example. The fact that it is possible to use HMMs that have been trained on cleanly recorded data, rich in phonetic contexts, as the basis for adaptation means that high-quality speech can be synthesised even when the adaptation data is noisy and sparse. Here, we present the application of speaker adaptation methods to the adaptation of adult-trained models to a child speaker for speech synthesis.

It should be noted also that speaker normalisation techniques figure prominently in work on using ASR systems trained on adult speech to recognise the speech of children. The use of Vocal Tract Length Normalisation is widely reported [7, 6, 4, 10, 11], and [11] also uses uniform scaling

of speaking rate. Voice conversion approaches that might be considered the synthesis equivalent of ASR speaker normalisation techniques could be used to convert the output of an adult-trained synthesiser to the voice of a child target speaker. But as successful alteration of segmental duration in voice conversion is not straightforward, it would be very difficult to achieve childlike hesitation or disfluency using these methods.

Dealing with missing units is straightforward in HMM-based voices, both speaker-dependent and speaker-adapted. HMM-based speech synthesis is able to construct a model for any missing unit, by sharing its parameters with existing models. This is achieved using similar data-driven, tree-based state clustering techniques to those used in ASR. In contrast, concatenative systems, when synthesising a sentence that requires an unseen unit (e.g., a particular diphone, or a diphone in a certain context), must select a substitute unit, typically on the basis of heuristics.

Together with the robustness of HMM-based speech synthesis to imperfect recording conditions [14], these capabilities are well suited to the task of child speech synthesis. We report an experiment comparing several configurations of an HMM-based speech synthesiser for child speech. We compared speaker-dependent and speaker-adaptive modelling for varying amounts of data in a listening test which evaluated similarity to the target speaker, naturalness and intelligibility. Since there may also be challenges in F0 tracking, spectral estimation and vocoding of child speech, we also evaluated vocoded speech alongside the synthetic speech.

# 2. BUILDING THE SYNTHESISER

## 2.1 Data Collection and Preparation

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 processing was slightly more complex than for adult data. The data were split into shorter fragments in order to exclude disfluencies, screaming, singing, sighs, page turns, and other non-speech sounds. We did not attempt to incorporate these elements into the synthetic voice. The data were hand-transcribed in standard orthography. Special care was taken to deal with mispronunciations and word-fragments in such a way that the final phonetic transcription would accurately reflect the contents of the audio files. Where there was a word in the lexicon that matched the speaker's mispronunciation, this word was used in the transcription (e.g. the speaker often read "cells" as "seals", and so the second word was used in the transcription). Where there was no existing lexical item to match the speaker's pronunciation of a word or fragment, an invented word was used in the normal spelling transcription, and then this invented word was added to the lexicon with the speaker's pronunciation before the phonetic transcription was generated.

At this stage, 30 sentences were chosen for their fair degree of fluency and medium length (4–9 words) from across the recording sessions and held out for use in evaluation.

A phone transcription was produced for the rest of the data with the Multisyn voice-building tools [2]. An initial phone transcription was produced by performing lexical look-up from the augmented lexicon. This initial transcription was then refined by forced alignment with the audio, in which vowel reduction and the insertion of pauses between words are allowed where supported by the audio data. Pause insertion is particularly important in the case of such hesitantly read data.

Three datasets (for training or adapting the HMMs) of varying size were constructed (small: 15 minutes of speech material; medium: 30 minutes; whole: all 94 minutes). Sentence order was randomised before partition into these three sets to avoid effects of the considerable difference between the recordings from different sessions.

## 2.2 HTS Systems Used

Two types of HMM-based speech synthesiser were built using HTS version 2.1 [17]: speaker-dependent and speaker-adaptive. The procedure used for building the speaker-dependent voices was the same as that for the HTS entry in the Blizzard Challenge 2005 [18]. The speaker-adaptive system adopted the gender-mixed average voice from the HTS entry in the Blizzard Challenge 2007 (using feature vectors with 40 mel-cepstral coefficients) [16]. Adaptation to the target speaker was performed with the procedure used for the same HTS entry in the Blizzard Challenge. A brief account of these procedures is given below.

## 2.3 Parameter Extraction

For both types of system built, the speech 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, to yield a 138-dimension observation vector for the HMMs. F0 was extracted using a three-stage procedure. First, the ESPS `get_f0` tool was used to extract F0 for all the speech data. These preliminary F0 values were then plotted as a histogram, from which a rough F0 range for the speaker was determined. F0 values were then re-extracted within the determined range using a voting method based on `get_f0`, `Tempo` and `IFAS`, the final F0 for each frame being the median of the three extracted values for that frame. Spectral analysis was performed with the high quality vocoder STRAIGHT [8], and the STRAIGHT spectra were converted to mel-cepstral coefficients.

## 2.4 Model Structure and Context Clustering

For both types of system built, speech units were modelled with HMMs of 5 emitting states in a left-to-right topology. In all cases, the same set of units was used: phones dependent not only on neighbouring phones, but on an extensive list of phonetic, linguistic and prosodic contexts (see [19] for the list).

The rich context-dependency of the speech units results in a very large number of models. This in turn means that almost all models will be sparsely represented in the training data (typically we find just one example of each in the training data!) and that, at synthesis time, models of missing units will certainly need to be created. Both of these problems are solved by the use of decision-trees. In the construction of these trees during training, model parameters are pooled and then repeatedly divided by the application of yes-no questions relating to the contextual features that define the models (e.g. 'Is the state part of a nasal consonant phone?',

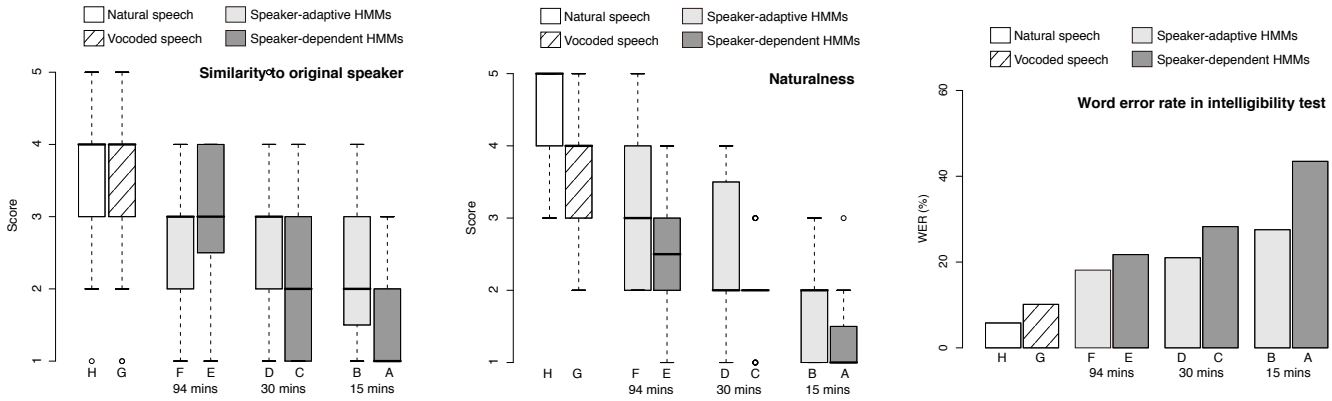


**Figure 1: Listening test results. Boxplot format follows [1]: "the median is represented by a solid bar across a box showing the quartiles; whiskers extend to 1.5 times the inter-quartile range and outliers beyond this are represented as circles."**

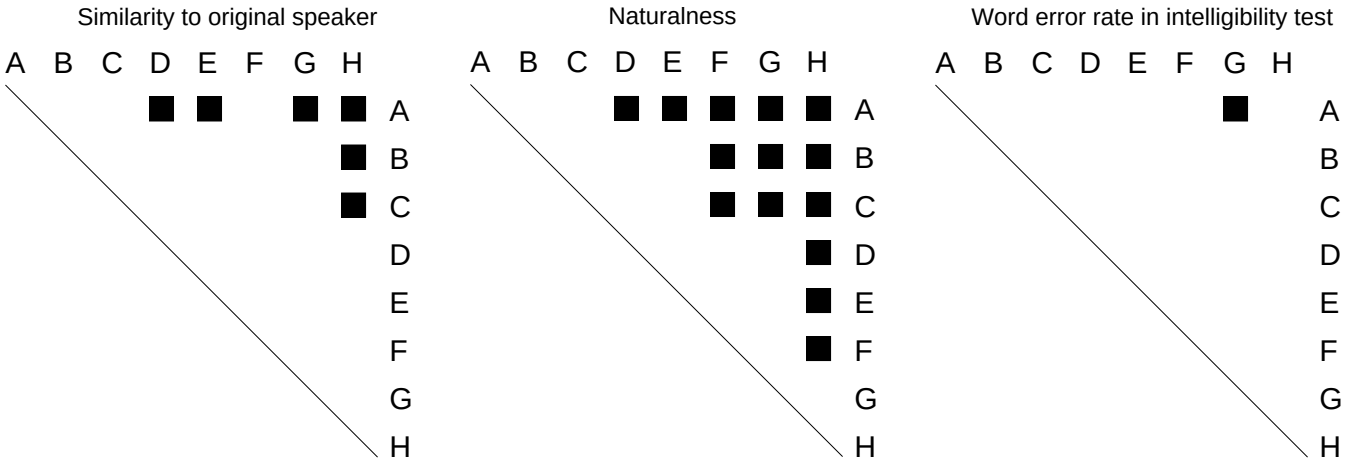


**Figure 2: Results of pairwise Wilcoxon signed rank tests between systems; a black square shows a significant difference between systems with $\alpha = 0.01$ (with Bonferroni correction).**

'Is the state part of a phone that occurs at the end of a word?' etc.). Questions are selected and ordered in the trees during training so that acoustically similar states end up pooled in the same leaf nodes of the trees. This solves the problem of data sparsity during training by allowing the parameters of acoustically similar states in a leaf node of the tree to be "tied" (re-estimated as a single distribution with the pooled training data). The trees solve the problem of unseen models at synthesis time by allowing the creation of these models: for each state of an unseen model, the relevant trees are traversed by answering the questions appropriately until a leaf node is reached. The probability distributions pointed to by this leaf node are then used to populate the relevant state of the unseen model.

Separate trees are made for spectral, F0 and aperiodicity measure distributions of each emitting state, and a single tree for duration is made for all states, resulting in 16 trees in the present set-up. This allows the clustering of units for spectral quality, F0, duration and aperiodicity measures with different trees using different context questions; as we would expect, different aspects of context affect the spectral quality than those affecting F0.

Although tree-building starts with a set of contexts (or yes-no context questions) which are hand-crafted to specify the phonetic and linguistic contexts which we think will have an effect on the acoustics of speech units in a given language, tree-building itself proceeds automatically. That is, questions are selected one by one according to some criterion and added to the tree until a stopping condition is met. In the current procedure, nodes associated with context questions are added to the trees until the MDL (Minimum Description Length)/BIC (Bayesian Information Criterion) criterion is met. The MDL/BIC criterion is a well-known information criterion for avoiding over-fitting of models to the training data and can specify an appropriate size for the decision tree [13].

## 2.5 Speaker Dependent System

For the speaker-dependent systems, model training began with the estimation of monophone models (phone models independent of context). These were then used as the basis for full-context models, which were re-estimated before decision-tree based context clustering was applied to spectral, log F0, aperiodicity and duration features separately. The clustered parameters were tied and re-estimated, then

**Table 1: Identifying letter for each system**

| Synthetic speech | | |
|---|---|---|
| | Modelling technique | |
| Amount of data from target speaker | Speaker dependent | Average voice adapted |
| 15 minutes | A | B |
| 30 minutes | C | D |
| 94 minutes | E | F |

| Natural speech | |
|---|---|
| Vocoded | G |
| Original | H |

the procedure was repeated: parameters were untied and re-estimated, clustered and and re-estimated a second time.

## 2.6 Speaker Adaptive System

As noted above, the speaker-adaptive system adopted an already-trained gender-mixed average voice from previous work. Details of training are given in [16]. This gender-mixed average voice model was trained on the six adult speakers of CMU-ARCTIC speech database (four male, two female). First, 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.

Adaptation of the gender-mixed average voice model was performed using data from the target speaker, the labels being modified to include target speaker gender. Adaptation was performed with a combination of constrained structural maximum a posteriori linear regression (CSMAPLR) and maximum a posteriori (MAP) adaptation.

## 2.7 Synthesis

The held-out sentences to be used in evaluation were synthesised with Festival. 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 models that had been trained, and waveforms were synthesised from those parameters.

# 3. EXPERIMENTS

## 3.1 Experimental Procedure

The evaluation of the various systems was carried out using a similar protocol to the Blizzard Challenge [5]. Each variant of our system was a 'participant' in this challenge. Included in the set of participants were two benchmarks – natural speech and vocoded natural speech – in which held-out sentences from the corpus were used, instead of actual synthesis. The vocoded speech was constructed by performing the speech analysis described earlier, followed by waveform generation without any modification of the features. The vocoder we use (STRAIGHT for analysis and a mixed-excitation source-filter model for waveform generation) does degrade the signal slightly, and we wished to evaluate the effect of this on child speech. The higher F0 value and higher formant frequencies of child speech, compared to adult speech, may cause spectral envelope estimation to be less accurate.

**Table 2: Sentences used to evaluate intelligibility**

| No. | Sentence text |
|---|---|
| 1 | I will eat only the pieces that fall off. |
| 2 | They rode away in trucks. |
| 3 | Snow? Mrs. Tate looked shocked. |
| 4 | Almost like diamonds, she said. |
| 5 | I am not a sheep, he said. |
| 6 | He put some salt on it. |
| 7 | The fire grew bigger. |
| 8 | He ran after little cats. |

The listening test, which was conducted via a web browser under quiet laboratory conditions using headphones, consisted of 3 sections. An 8-by-8 Latin Squares design was employed. There were 8 systems and 8 listener groups, with 8 different utterances per section (a total of 24 different utterances). In any given section, a single listener group heard every system once, each time with a different utterance. Every system was used to synthesise every utterance once within each section. We used a total of 24 paid listeners (3 people per listener group), who were all native speakers of English between the ages of 18 and 25.

In the first section, listeners were asked to rate the similarity of each stimulus to the original speaker. Two natural reference utterances were provided, which listeners could play at any time, as many times as they wished. Listeners could also listen to each stimulus as many times as they wished. A five point scale was used; the end points of the scale were described to the listeners as "1 – Sounds like a totally different person" and "5 – Sounds like exactly the same person".

The second section followed the same format as the first, but this time listeners were asked to rate the naturalness of each stimulus on a 5 point scale, with end points described to the listeners as "1 – Completely Unnatural" and "5 – Completely Natural".

In the final section, listeners were asked to type in a transcription of each test stimulus. Normally, we would use Semantically Unpredictable Sentences for this type of test, to avoid ceiling effects on transcription accuracy. However, we felt that such sentences sounded extremely unnatural when uttered by a synthetic child voice. Additionally, we did not have natural recordings of the speaker saying such sentences. Therefore, we used sentences held out from the corpus for this part of the test. These sentences are listed in Table 2.

## 3.2 Results and Discussion

The listening test data were analysed using the same statistical techniques used in the Blizzard Challenge 2007 [1], and we present results in Figure 1. Significant differences

between systems are presented in Figure 2. The differences in the results for all three sections are measured by the same test used in the Blizzard Challenge 2007: a Wilcoxon signed rank test with $\alpha = 0.01$ and Bonferroni correction. It should be noted that WER was computed from a set of sentences of differing lengths; six of the sentences consist of 4–6 words and the remaining two, 7 and 9 words. This was necessitated by the fact that the test sentences were naturally occurring sentences, 'harvested' from the recordings rather than generated specifically for the evaluation. This had an unfortunate consequence: the within-subjects design of the Wilcoxon test used meant that significant differences between systems for WER had to be based on scores for each listener for each system already normalised for word length. However, it was not thought that the sentences vary greatly enough in length that the outcome of the significance test for WER would be seriously affected by this.

There are several trends observable in Figure 1 which receive partial or no support from the significance test, but which we expect would be detected as significant effects under more extensive evaluation, with a greater number of listeners giving greater statistical power.

In most cases increasing the amount of training or adaptation data gives a higher median score in sections 1 and 2 and a lower mean WER in section 3 between systems of the same type, as we would expect. In three cases this effect was found to be significant (between systems A and E in section 1 and between systems A and E and systems B and F in section 2), and we would expect to find a greater number of significant differences if a more extensive listening test were to be performed. We note that the Blizzard Challenge uses many hundreds of listeners, yet still cannot detect statistically significant differences between all pairs of participating systems.

In most cases, a speaker-adaptive voice yields higher median opinion scores and lower mean WER than a speaker-dependent voice trained on the same amount of data. Although none of these differences were found to be statistically significant, this is a trend that we would expect in the light of previous research showing that adaptation of an average voice with a few minutes of target speaker data results in more natural synthetic speech than the training from scratch of a speaker dependent voice on a larger dataset [16]. It should be noted that in the present case, the average voice was trained on very different speakers (adults) to our target speaker (a child), and yet the same result appears to hold. Despite speaker differences, the average voice – trained on plentiful context-rich data – nevertheless incorporates a lot of prior knowledge about speech in general and can provide the basis for successful speaker adaptation.

There is an interesting exception to the two trends mentioned above in the case of section 1 of the evaluation. When the amount of data is increased from 30 to 94 minutes in this section, the median similarity of the speaker-dependent voice to the original speaker increases but the median for the average voice-based system remains unchanged. The median similarity score of the speaker-dependent voice is the same as that of the adapted voice. This suggests that improvements in similarity to the original speaker achieved by increasing the size of the dataset are smaller when performing adaptation than when training speaker-dependent voices. Similarity to the original speaker is perhaps the aspect of the speaker-adaptive approach that needs the most improvement.

In the evaluation of naturalness, the natural vocoded speech received a median opinion score of one point less than that of the original speech, and in the evaluation for intelligibility, it received a higher mean WER. Although neither of these differences was found to be statistically significant, these scores suggest that vocoding alone is causing degradation of the speech signal. The difference in scores was larger than we had expected; whether this degradation in quality is specific to child speech could be the subject of useful future research.

In the evaluation of similarity to the original speaker, even the natural speech received a median opinion score of 4, where we would expect "5 – Sounds like exactly the same person". This might be attributed to the variability of the child speech data: the two natural speech samples given for reference in the evaluation were taken from different recording sessions and have slightly different qualities. The synthetic speech in effect "averages out" the speaker/recording condition variability across all the data, and as such is different in quality from either of the two natural samples. If it were possible to evaluate voices built on more consistent data, then we would expect natural speech to receive a median opinion score of 5 in this section of the evaluation.

# 4. CONCLUSIONS

This paper has described the application of existing HMM-based speech systems to the synthesis of a child's speech. We have built both speaker-dependent voices and speaker-adapted voices, where an average voice model which had previously been trained on adult speakers was successfully adapted to the child target speaker.

We consider the experiment successful in that the synthetic speech clearly reflects the qualities of the training data. That is, the synthetic speech sounds like the speech of a child reading, with the same patterns of hesitancy and disfluency that we observed in the training data. However, such disfluency will not be desirable in all applications of such a synthetic voice, and ways of synthesising fluently-spoken child speech without recording additional data will be the subject of future work. One way to achieve this would be to combine models trained on different speakers. Previous work has examined the interpolation of different emotions and speaking styles [12]; it would be interesting to hear the results of interpolation of speakers of different ages. Another method would be to use a combination of F0, spectral and duration models taken from different speakers. For example, a model composed of the duration model from a fluent adult speaker and the spectral and F0 models of a child speaker may result in fluent yet child-like speech. We plan to try this in future.

Examples are available at `http://homepages.inf.ed.ac.uk/s0676515/child_speech`

# 5. ACKNOWLEDGEMENTS

The authors would like to thank Robyn Zundel, daughter of Kay Berkling, for her contribution to this research through many hours of reading and recording. It was not easy work, but she did it for the research!

# 6. REFERENCES

[1] R. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King. Statistical analysis of the Blizzard Challenge 2007 listening

test results. In *Proc. BLZ3-2007 (in Proc. SSW6)*, Aug. 2007.

[2] R. A. J. Clark, K. Richmond, and S. King. Multisyn: Open-domain unit selection for the Festival speech synthesis system. *Speech Communication*, 49(4):317–330, 2007.

[3] P. Cosi and L. Bryan. Italian children's speech recognition for advanced interactive literacy. In *Proc. Interspeech 2005*, pages 2201–2204, Sept. 2005.

[4] S. Das, D. Nix, and M. Pichen. Improvements in children's speech recognition performance. In *Proc. ICASSP-98*, pages 433–436, May 1998.

[5] M. Fraser and S. King. The Blizzard Challenge 2007. In *Proc. BLZ3-2007 (in Proc. SSW6)*, Aug. 2007.

[6] D. Giuliani and M. Gerosa. Investigating recognition of children's speech. In *Proc. ICASSP 2003*, pages 137–140, Apr. 2003.

[7] J. Gustafson and K. Sjölander. Voice transformations for improving children's speech recognition in a publicly available dialogue system. In *Proc. ICSLP 2002*, pages 297–300, Sept. 2002.

[8] 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*, 27:187–207, 1999.

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

[10] A. Potamianos and S. Narayanan. Robust recognition of children's speech. *IEEE Trans. on Speech Audio Process.*, 11(6):603–616, Nov. 2003.

[11] G. Stemmer, C. Hacker, S. Steidl, and E. Nöth. Acoustic normalization of children's speech. In *Proc. EUROSPEECH 2003*, pages 1313–1316, Sept. 2003.

[12] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi. Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing,. *IEICE Trans. Inf. & Syst.*, E88-D(11):2484–2491, Nov. 2005.

[13] 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. *IEEE Trans. Speech, Audio & Language Process.*, 2007. (accept).

[14] J. Yamagishi, Z. Ling, and S. King. Robustness of HMM-based speech synthesis. In *Interspeech 2008 (accepted)*, Sept. 2008.

[15] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi. A training method of average voice model for HMM-based speech synthesis. *IEICE Trans. Fundamentals*, E86-A(8):1956–1963, Aug. 2003.

[16] 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.

[17] 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*, pages 294–299, Aug. 2007.

[18] H. Zen, T. Toda, M. Nakamura, and K. Tokuda. Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005. *IEICE Trans. Inf. & Syst.*, E90-D(1):325–333, Jan. 2007.

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