Synthesis of Child Speech With HMM Adaptation and Voice Conversion

Oliver Watts, Junichi Yamagishi, *Member, IEEE*, Simon King, *Senior Member, IEEE*, and Kay Berkling, *Senior Member, IEEE*

***Abstract*—The synthesis of child speech presents challenges both in the collection of data and in the building of a synthesizer from that data. We chose to build a statistical parametric synthesizer using the hidden Markov model (HMM)-based system HTS, as this technique has previously been shown to perform well for limited amounts of data, and for data collected under imperfect conditions. Six different configurations of the synthesizer were compared, using both speaker-dependent and speaker-adaptive modeling techniques, and using varying amounts of data. For comparison with HMM adaptation, techniques from voice conversion were used to transform existing synthesizers to the characteristics of the target speaker. Speaker-adaptive voices generally outperformed child speaker-dependent voices in the evaluation. HMM adaptation outperformed voice conversion style techniques when using the full target speaker corpus; with fewer adaptation data, however, no significant listener preference for either HMM adaptation or voice conversion methods was found.**

***Index Terms*—Children, hidden Markov models (HMMs), speech synthesis.**

I. Introduction

THE synthesis of child speech presents special difficulties for data-driven speech synthesis systems due to the type of child speech corpus typically available. Unit selection speech synthesis (e.g., [3]), which has come to be the dominant approach to data-driven speech synthesis over the last decade, produces waveforms for arbitrary novel utterances by directly reusing existing sections of waveform from a database. One of the major strengths of this approach is that in ideal conditions, natural waveforms are reused directly with no need for manipulation of spectrum or fundamental frequency and the degradation of speech quality that this manipulation entails. However, some of the major drawbacks of unit selection stem from the same source: if the database is imperfect, this will have a direct impact on the quality of the speech synthesized. Imperfections include speaker inconsistency, background noise, and poor phonetic coverage, all problems typically associated with available child speech data.

Statistical parametric approaches to speech synthesis (such as hidden Markov model (HMM)-based speech synthesis) have grown in popularity over the last few years [4]. These approaches have been shown to be less sensitive than unit selection to imperfect training data [5]. In HMM-based speech synthesis, use of parameter sharing techniques allows the synthesis of models for speech units unseen in the training corpus; this contrasts with the corresponding strategy that must be used in unit selection where the system must select a substitute unit, typically on the basis of heuristics. Unlike unit selection synthesis, already-trained HMM-based systems can be adapted to the voice characteristics of a target speaker with small amounts of adaptation data. 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 synthesized even when the adaptation data is noisy and sparse.

Adaptation has been used to impose various types of characteristics onto existing statistical parametric synthesizers, for example, characteristics associated with dialect [6] and speaking style [7]. In [1], we applied adaptation techniques (among others) to the creation of what (to our knowledge) is the first data-driven synthesizer of child speech. We present this work here with fuller analysis and with a comparison between HMM adaptation techniques and techniques from voice conversion for the transformation of an existing synthesizer to a child speaker.

This paper is organized as follows. Section II describes a corpus of child speech data collected especially for this research and compares its suitability for use in training speech synthesizers with that of a purpose-built speech synthesis corpus. Section III describes the systems built with the child speaker as target speaker, and Section IV reports evaluation of these systems. We conclude the paper by briefly summarizing our findings in Section V.

II. Child Speech Data

The training of data-driven synthesis systems for child voices presents difficulties due to the type of data which it is generally feasible to collect from child speakers. Data-driven speech synthesizers are conventionally trained on corpora that are phonetically balanced, consistently read, and cleanly recorded. A

Manuscript received May 11, 2009; revised September 22, 2009. Current version published June 16, 2010. This work was supported in part by the European Community's Seventh Framework Program (FP7/2007-2013) under Grant 213845 (the EMIME project) and in part by the Edinburgh Compute and Data Facility which is partially supported by the eDIKT initiative (http://www.edikt.org.uk). The work of J. Yamagishi was supported by the Engineering and Physical Sciences Research Council (EPSRC) and EMIME. The work of S. King was supported by an EPSRC Advanced Research Fellowship. Part of the work reported here was described in [1] and part was introduced in [2] . The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Chung-Hsien Wu.

O. Watts, J. Yamagishi, and S. King are with the Centre for Speech Technology Research, University of Edinburgh, Edinburgh EH8 9YL, U.K. (e-mail: O.S.Watts@sms.ed.ac.uk, jyamagis@inf.ed.ac.uk, simon.king@ed.ac.uk).

K. Berkling is with Inline Internet Online Dienste GmbH, 76133 Karlsruhe, Germany (e-mail: kay@berkling.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2009.2035029

1558-7916/$26.00 © 2010 IEEE

good example of such a corpus is the CMU Arctic database, designed and recorded by several adult speakers specifically for the purpose of speech synthesis [8]. The type of child speech typically available, on the other hand, more closely resembles "found" data in that it does not give good coverage of the phonetic/prosodic units of the language, is inconsistently read, and is imperfectly recorded. An example of such a child speech corpus is the one collected and prepared in the work described here. A comparison of this corpus with one part of the CMU Arctic database (the data of speaker SLT, which will here be referred to simply as *SLT*) will be made in order to highlight the problems inherent in the sort of child speech data corpus typically available. We begin with a brief overview of the collection of what will here be called the *Child* corpus before moving on to make the comparison with the more conventional database.

A. Child Database: Overview

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 parts of the story book text, which she was allowed to read without interruption. A total of just over 100 minutes (after processing) of speech data were collected.

B. Recording Conditions

A database to be used for speech synthesis should ideally be well recorded, free from reverberation and background noise and have consistent acoustic quality. This is the case with *SLT*, which was recorded in a purpose-built studio. As noted above, the *Child* corpus—on the other hand—was recorded using a headset microphone in the home of one of the authors: it is more difficult to get a child into the studio than a paid voice talent. Consequently, the recordings contain considerable background noise, including the sounds of traffic and wildlife, and reverberation. The speech was collected without interruption and so utterances were interspersed with sighs, page turns, and other non-speech sounds which it was not possible in all instances to remove without also excising valuable speech data. Fig. 1 shows spectrograms of three short excerpts from the recordings showing background noise fairly typical of these data.

C. Speaker Variability

Children do not typically have the vocal and emotional control necessary to minimize inconsistency during and between recording sessions. Use of a coherent "script" (storybook text) 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 which can be heard in the SLT data, with the child speaker screaming, singing, and chanting at appropriate points in the story and deliberately altering voice quality to act out characters' parts. Children can only be persuaded to record data in short sessions, with the result that the recordings of the *Child* corpus were made over several months, an additional cause of inter-session variation.

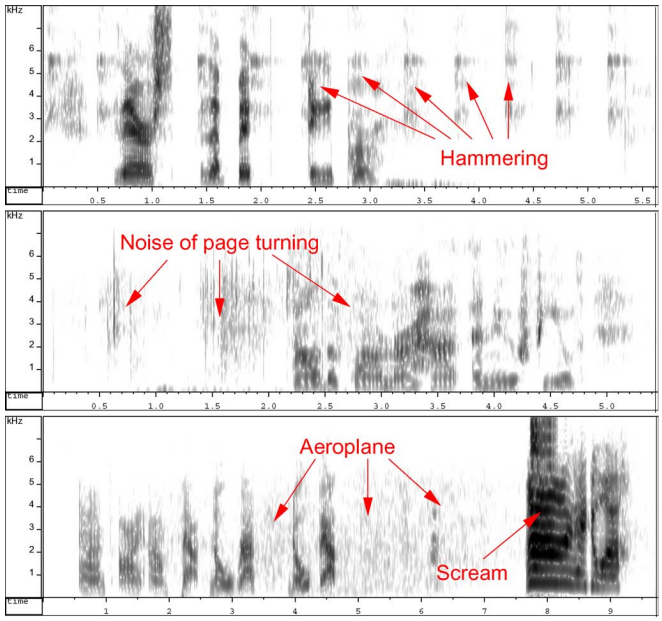

Fig. 1. Spectrograms of portions of the collected child data, showing overlapping speech and background noise. The North American-accented English speech of a seven-year old trilingual (Spanish, English, German) female was collected using a headset microphone in an informal setting at the home of one of the authors over the course of several months. Consequently, the recordings contain considerable nonspeech sounds, including the sounds of hammering, page turning, traffic, and wildlife, and reverberation.

Artistic engagement with the recording script and short recording sessions aside, a high degree of innate variability is a well-known characteristic of children's speech, in comparison with the speech of adults (e.g., [9]). Fig. 2 plots of F_0 values, power, and duration of a single phone type (/aa/) from the *Child* and *SLT* databases. It can be seen that not only does the child speech have generally higher fundamental frequency values, lower power, and longer duration values than that of the adult, but all these factors have wider ranges.

D. Database Coverage

An established way to create prompts for speech synthesis databases is to select utterances from a large corpus of text according to some criterion of phonetic coverage (e.g., [8]). Recording scripts resulting from this type of procedure, while phonetically well-balanced, are not coherent texts that a seven-year-old child could be persuaded to read. It is more feasible to use, for example, story books familiar to the child, as was done in the work presented here. This results in a corpus with comparatively poor phonetic coverage. Table I gives figures for triphone and quinphone coverage of four corpora. The information given for each corpus includes number of tokens, number of triphone and quinphone types, along with Type-Token Ratios (TTR) for triphones and quinphones. TTRs are a measure of the richness of the content of the corpora, a TTR of 1.0 indicating that a each token of a corpus is of a different type, thus representing the maximum coverage a corpus of that size could provide. The first two corpora are the *Child* corpus and *SLT*. *SLT2* denotes a large subset taken from the beginning of the SLT corpus which contains the same number of tokens as the *Child* corpus. This

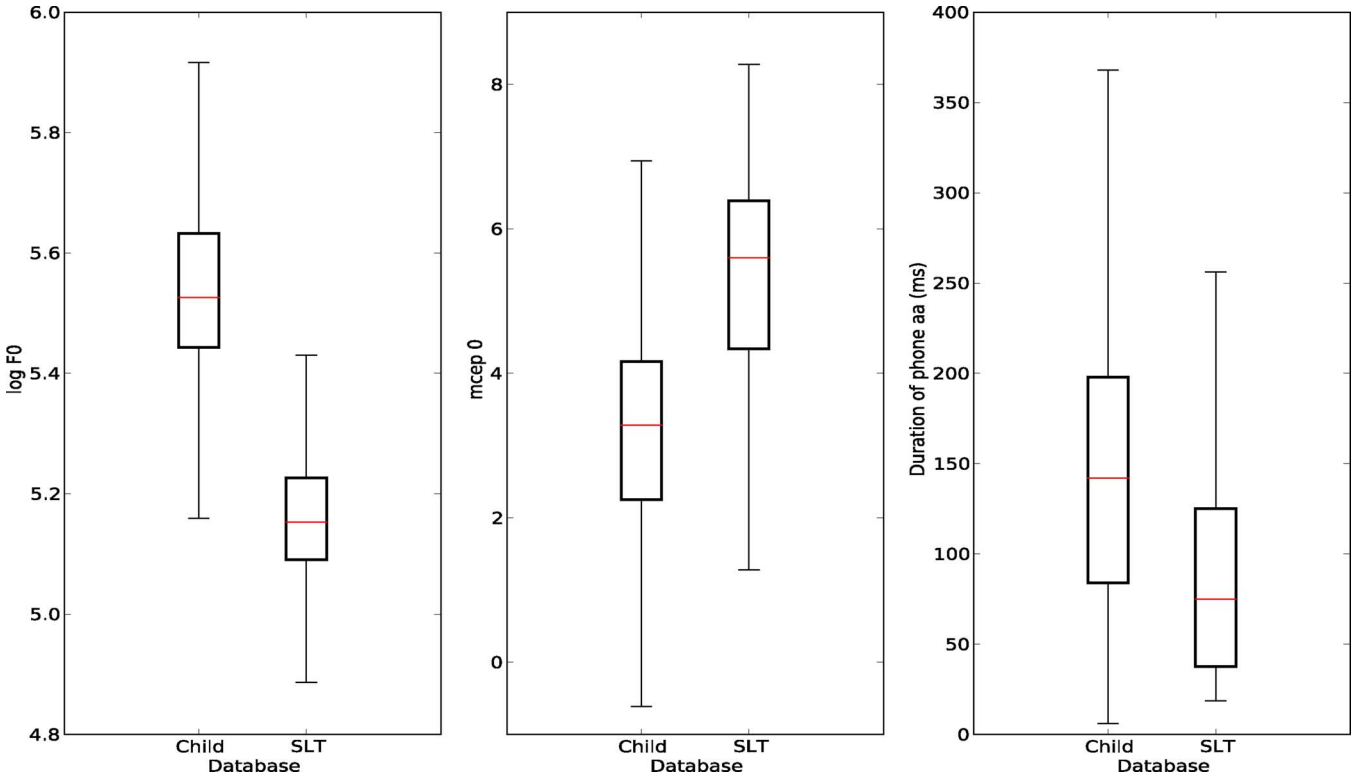

Fig. 2. Plots of F_0, power and duration of phone /aa/ for *Child* and *SLT* corpora. Medians are shown as red bars across boxes indicating quartiles, and whiskers extend to 1.5 times the inter-quartile range. It can be seen that not only does the child speech have generally higher fundamental frequency values, lower power, and longer duration values than that of the adult, but all these factors have wider ranges.

TABLE I

COVERAGE OF VARIOUS DATABASES. SLT WAS COMPILED FOR SPEECH SYNTHESIS USING A PHONETIC COVERAGE CRITERION, SLT2 IS A SUBSET OF SLT OF THE SAME SIZE AS CHILD INCLUDED HERE FOR FAIR COMPARISON OF TYPE-TOKEN RATIOS. SNARK WAS COMPILED FROM TEXT OF THE SAME DOMAIN AS SLT BUT WITHOUT THE PHONETIC COVERAGE CRITERION. THE TOP HALF GIVES COUNTS AND RATIOS OVER ALL UNITS IN THE CORPORA. THE BOTTOM HALF GIVES INFORMATION EXCLUDING TOKENS CONTAINING THE *PAU* TOKEN (EITHER AS CENTRAL PHONE OR ANY OF THE FOUR POSSIBLE LEFT AND RIGHT CONTEXTS). THE CHILD CORPUS COLLECTED USING CHILDREN'S STORIES HAS WORSE COVERAGE NOT ONLY THAN THE PURPOSE-BUILT RECORDING SCRIPT, BUT ALSO THAN NORMAL LITERATURE FOR ADULTS WITH NO SELECTION CRITERION APPLIED. (a) ALL TYPES/TOKENS COUNTED. (b) TYPES/TOKENS CONTAINING PAUSE NOT COUNTED

(a)

Database	Child	SLT	SLT2	Snark
Tokens	37707	38866	37707	37707
Triphone types (*TTR*)	6376 (*0.17*)	9546 (*0.25*)	9430 (*0.25*)	8165 (*0.22*)
Quinphone types (*TTR*)	20478 (*0.54*)	28662 (*0.74*)	27912 (*0.74*)	26601 (*0.71*)

(b)

Database	Child	SLT	SLT2	Snark
Tokens	18555	30377	18555	18555
Triphone types (*TTR*)	4773 (*0.26*)	8482 (*0.28*)	6676 (*0.36*)	5775 (*0.31*)
Quinphone types (*TTR*)	11246 (*0.61*)	24035 (*0.79*)	15665 (*0.84*)	14531 (*0.78*)

subset was included because corpus size influences the magnitude of type-token ratios; a subset of the same length, however, allows fair comparison of TTRs. Finally, *Snark* is a corpus compiled especially for this analysis. A story (Jack London's *Voyage of the Snark*) from the same domain as the Arctic texts (and in fact included among them) was split into sentences, which in turn were turned into linguistic specifications with the front end of the Festival synthesizer and finally context-dependent labels. A subset was taken from the beginning having the same number of tokens as the *Child* corpus. The *Snark* corpus was assembled to provide a midway point for comparison: the fact that it was taken from the same domain as Arctic but that utterances were taken sequentially from the story rather than being selected according to a criterion of phonetic coverage means that the figures for this corpus allow us to assess in isolation the influence of the selection procedure on the coverage of the resulting corpora.

The top half of Table I gives counts and ratios over all units in the corpora. It was noted that *Child* contains a much greater number of pauses (represented by the "phone" *pau*) than the other databases, including a great number of consecutive pause tokens due to relatively long hesitations and disfluencies. Although these pauses are one of the factors which strongly char-

acterize the *Child* corpus, it was thought informative to compute similar information, discounting tokens containing the *pau* token (either as central phone or any of the four possible left and right contexts). This information is given in the bottom half of the table.

The coverage statistics computed here (triphones and quinphones) reflect some of the contextual factors considered during HMM-based voice building (see Section III below). There are a great many other phonetic, prosodic, and syntactic factors considered during voice building that have been ignored in this analysis. Note also that although the criterion used in the construction of the SLT corpus (diphone coverage) is related only indirectly to triphone and quinphone coverage, the triphone and quinphone coverage are improved by prompt selection using the diphone criterion. That is, both when pauses are counted and when they are ignored, for both triphone and quinphone, *SLT2*'s TTRs are higher than *Snark*'s. The *Child* database has lower TTRs than those of either *SLT2* or *Snark* in all cases. That is, the corpus collected using children's stories has worse coverage not only than the purpose-built recording script, but also than normal literature for adults with no selection criterion applied.

Although monophone coverage is not shown in Table I, it should be noted that all corpora studied give complete coverage of the set of monophones except *Child*, where one phone of the phoneset used (/zh/) is absent entirely from the corpus.

Table II lists the ten most common quinphones (excluding quinphones containing the *pau* token) of *Child*, *SLT2*, and *Snark*, which give clues about the reasons for the worse coverage of *Child*. Eight of the *SLT2* quinphones come from sequences of function words or function words and verbs ("it was," "there was," "he was," "he had," etc.); only 1 comes from a person's name ("Phillip"). On the other hand, seven of the *Child* quinphones come from people's names ("Pickle," "Mrs.," "Dragon," "Christy," "Greg"), and only 1 quinphone from a function word-verb sequence ("he said") occurs among the most frequent ten. One of the reasons for *Child*'s relatively poor phonetic coverage, then, is its repetition of proper names. We note that repetitiveness is also seen more generally in the corpus. Note the following utterances from the corpus which exemplify the sort of repetition found extensively in these children's stories.

(1) She knew how to set up a tent, she knew how to build a camp fire, she knew how to cook camp food.
(2) The car began to roll. Faster and faster and faster and faster.

This section has outlined some of the differences between the target speaker data available to us and a conventional speech synthesis database. Despite the imperfect nature of the *Child* corpus, we wished to produce a text-to-speech system with the voice characteristics of the child target speaker. The measures taken to overcome the difficulties presented by the data consisted firstly of especially careful preparation of the data and secondly of choice of synthesis methods appropriate to this data. We conclude Section II with an account of the front-end processing involved in preparing the *Child* corpus before turning to synthesis methods employed in the following section.

TABLE II
MOST FREQUENT QUINPHONE TYPES IN THREE CORPORA. EIGHT OF THE *SLT2* QUINPHONES COME FROM SEQUENCES OF FUNCTION WORDS OR FUNCTION WORDS AND VERBS ("IT WAS," "THERE WAS," "HE WAS," "HE HAD," *ETC.*); ONLY ONE COMES FROM A PERSON'S NAME ("PHILLIP"). SEVEN OF THE *CHILD* QUINPHONES COME FROM PEOPLE'S NAMES ("PICKLE," "MRS.," "DRAGON," "CHRISTY," "GREG"), AND ONLY ONE QUINPHONE FROM A FUNCTION WORD–VERB SEQUENCE ("HE SAID") OCCURS AMONG THE MOST FREQUENT TEN

	SLT2		Child		Snark	
Rank	#	Type	#	Type	#	Type
1	31	ih t w aa z	76	p ih k ax l	28	dh ax s n aa
2	23	eh r w aa z	75	ih k ax l z	27	s n aa r k
3	21	dh eh r w aa	51	m ih s ax s	27	ax s n aa r
4	19	hh iy w aa z	50	r ae g ax n	23	ae n d dh ax
5	19	hh iy hh ae d	50	d r ae g ax	18	ih t w aa z
6	19	f ih l ax p	38	l ih t ax l	18	b iy f ao r
7	16	ih n hh ih z	38	k r ih s t	14	f ao r dh ax
8	14	hh ae d b ih	37	d g r eh g	13	n ae v ax g
9	13	s eh l v z	36	hh iy s eh d	13	ae v ax g ey
10	13	l r eh d iy	36	f aa r m er	11	w aa z dh ax

E. Data Preparation

The data were recorded without interruption and so had to be split into shorter fragments in order to exclude disfluencies, screaming, singing, sighs, page turns, and other nonspeech 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. A phone transcription was produced for the rest of the data with the Multisyn voice-building tools [3]. 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.

III. THE SYSTEMS

A. Speaker Dependent and Speaker Adapted HTS Voices

HMM-based synthesis was our chosen method for building voices because of its robustness to imperfect recording conditions [5], its integrated data-driven method for synthesizing units missing from the training corpus, and the possibility of speaker adaptation which it offers. We initially built two types of HMM-based speech synthesizer—speaker dependent and

TABLE III
IDENTIFYING LETTERS USED FOR EACH SYSTEM. TRANSFORMATIONS ARE TO THE TARGET SPEAKER IN ALL CASES EXCEPT FOR THE DURATION ADAPTATION OF SYSTEM I. HMM: HMM ADAPTATION (CSMAPLR + MAP), VC: VOICE CONVERSION. FIRST GROUP (A–F) IS DESIGNED FOR COMPARISON OF SPEAKER-DEPENDENT AND SPEAKER-ADAPTIVE TECHNIQUES WITH THREE TARGET SPEAKER CORPUS SIZES, THE SECOND GROUP IS DESIGNED FOR COMPARISON OF VOICE CONVERTED UNIT SELECTION SYSTEMS WITH HMM-BASED ONES AND TO ASSESS THE IMPACT TRANSFORMING DURATION TO TARGET SPEAKER IN BOTH CASES, AND THE LAST TWO GROUPS ARE DESIGNED FOR MORE CONTROLLED COMPARISON OF HMM ADAPTATION AND VOICE CONVERSION METHODS WITH TWO SIZES OF TARGET SPEAKER CORPUS

	Text-to-speech					
	Base voice	Base voice	Target	Transformation method		
ID	Type	speaker(s)	data	Spec.	F_0	Dur.
A	HTS	Child	15 min	—	—	—
B	HTS	6 adults	15 min	HMM	HMM	HMM
C	HTS	Child	30 min	—	—	—
D	HTS	6 adults	30 min	HMM	HMM	HMM
E	HTS	Child	94 min	—	—	—
F	HTS	6 adults	94 min	HMM	HMM	HMM
I	HTS	6 adults	94 min	HMM	HMM	HMM (to SLT)
J	Unit sel.	SLT	94 min	VC	VC	—
K	Unit sel.	SLT	94 min	VC	VC	VC
L	HTS	SLT	15 min	HMM	HMM	HMM
M	HTS	SLT	15 min	VC	HMM	HMM
N	HTS	SLT	15 min	HMM	VC	HMM
O	HTS	SLT	15 min	HMM	HMM	VC
P	HTS	SLT	94 min	HMM	HMM	HMM
Q	HTS	SLT	94 min	VC	HMM	HMM
R	HTS	SLT	94 min	HMM	VC	HMM
S	HTS	SLT	94 min	HMM	HMM	VC
	Natural speech					
G	Vocoded natural speech					
H	Original speech					

speaker adaptive—using three different size subsets of the *Child* corpus as the target speaker corpus. These six systems (A–F) are listed at the top of Table III. The systems were built using HTS version 2.1 [10], and their construction is outlined here. To build Systems A, C, and E (speaker dependent) we followed the procedure used for the HTS entry in the Blizzard Challenge 2005 [11]. Systems B, D, and F (speaker adaptive) were based on a gender-mixed average voice from the HTS entry in the Blizzard Challenge 2007 [12]. Adaptation of this existing average voice to the child speaker data followed the procedure used for the same HTS entry in the Blizzard Challenge. An account of these procedures is given below, beginning with details of parameter extraction and model structure common to all six systems.

1) Parameter Extraction: For both types of system built, the speech was parameterized as 40 mel-cepstral coefficients, $\log F_0$ and the energy of aperiodic components in five frequency bands, and the dynamic and acceleration features derived from all of these, to yield a 138-dimension observation vector for the HMMs. F_0 was extracted using a three-stage procedure. First, the ESPS `get_f0` tool was used to extract F_0 for all the speech data. These preliminary F_0 values were then plotted as a histogram, from which a rough F_0 range for the speaker was determined. F_0 values were then re-extracted within the determined range using a voting method based on `get_f0`, `Tempo`, and `IFAS`, the final F_0 for each frame being the median of the three extracted values for that frame. Spectral analysis was performed with the high quality vocoder STRAIGHT [13], and the STRAIGHT spectra were converted to mel-cepstral coefficients.

2) Model Structure and Context Clustering: For both types of system built, speech units were modeled with HMMs of five emitting states in a left-to-right topology. For state emission probability density functions (pdfs), single mixture component Gaussian distributions with diagonal covariance matrices were used. In all cases, the same set of units was used: phones dependent not only on neighboring phones, but on an extensive list of phonetic, linguistic, and prosodic contexts (see [14] 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?", "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, F_0 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 setup. This allows the clustering of units for spectral quality, F_0, duration and aperiodicity measures with different trees using different context questions; as we would expect, different aspects of context affect spectral quality than those affecting F_0. Although tree-building starts with a set of contexts (or yes-no context questions) which are handcrafted 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 minimum description length (MDL) criterion is met. The MDL 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 [15].

3) Speaker-Dependent Systems (A, C, and E): 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, $\log F_0$, aperiodicity and duration features separately. The clustered parameters were tied and re-estimated, then the procedure was repeated: parameters were untied and re-estimated, clustered, and re-estimated a second time.

4) Speaker Adaptive Systems (B, D, and F): As noted above, the speaker-adaptive system adopted an already-trained gender-mixed average voice from previous work. Details of training are given in [12]. 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 normalization 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 normalization being determined from the gender-mixed decision-trees. State durations obtained during this estimation were used to initialize duration probability distributions which were then clustered. SAT was performed on the complete HSMMs to re-estimate all parameters (including duration) with speaker normalization.

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 [15].

5) Synthesis: During preparation of the *Child* corpus, 30 sentences had been chosen for their fair degree of fluency and medium length (4–9 words) from across the recording sessions and held out from training and adaptation. These utterances were synthesized with Festival. Festival's front-end performed the phonetic and linguistic predictions needed to provide 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 synthesized from those parameters.

B. Unit Selection Synthesis + Voice Conversion

We wished to explore the possibilities offered by voice conversion techniques for imposing a novel speaker's voice characteristics on an existing synthesizer and to compare them with HMM-Based methods. 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: [16] and [17]) have been applied to the conversion of speech synthesizers' output to the voice characteristics of new speakers. For example, in [18], the output of a concatenative (diphone) synthesizer is used to create "source speaker" training data for a voice conversion model. Arbitrary novel utterances subsequently produced by the synthesizer can then be converted to the voice characteristics of the target speaker.

The application of voice conversion to the output of an existing speech synthesizer is particularly attractive in the context of sample-based concatenative methods where there is no statistical model whose parameters can be transformed. We therefore sought to compare the voice-converted output of a unit selection synthesizer with the results of speaker adaptation of HMM-based systems.

In GMM-based voice conversion, source speaker's durations are typically used unconverted in the output speech; this is the case in our system J, where spectrum and F_0 of the synthesizer's output were converted. This can be successful in cases where the differences between the durational characteristics of source and target speakers are negligible (as might be true of a pair of adult speakers), but in the present case the durational patterns which characterize the target speaker are very different from those of the adult speaker on whose data the base voice was trained. It is unreasonable to think that reusing source speaker's durations unmodified could be successful in this case. Therefore, we built a second system K where in addition to performing spectral and F_0 conversions, converted utterances' durations were obtained by uniformly stretching by a predetermined factor the utterances which had been output by the synthesizer.

1) Training (systems J and K): The synthesizer used as the "source speaker" in both systems J and K (see Table III) was an existing unit selection voice which had been built with the *SLT* corpus, using the Multisyn voice building tools ([3]).

Voice conversion models to map from the output of this base synthesizer to the characteristics of the child target speaker were trained using scripts distributed as part of the FestVox project, implementing techniques developed by Toda ([17], [19]). Training a Gaussian mixture model (GMM) for spectral conversion requires a parallel corpus in which the source and target speakers each utter the same words. The missing "source speaker" half of this corpus was synthesized using the base unit selection synthesizer; the target speaker half was natural data consisting of the whole of our target speaker training data (94 min). At no point after the synthesis of this "source speaker" data was knowledge of the linguistic contents of the utterances used; in training the conversion models and applying those models to arbitrary new speech it was assumed that knowledge of phone, word, etc., alignment was unavailable.

The speech was parameterized using the STRAIGHT mel-cepstral analysis and F_0 extraction described in Section III-A1 above rather than with the analysis tools distributed with the FestVox module. The only departure from the analysis procedure described in Section III-A1 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 initialized using vector quantization, and then iteratively refined using expectation–maximization. The data alignment and GMM training were iterated.

The conversion model for F_0 was obtained by computing the mean and standard deviation of both source and target speakers' $\log F_0$. Additionally, for converting duration in system J, a duration scaling factor was computed as the ratio of the total duration of source speaker training data to the duration of that of the target speaker.

2) Synthesis and Conversion: The sentences to be used in evaluation were taken from the story *Goldilocks and the Three Bears*; they were synthesized with Festival's front-end as in Section III-A5, 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 analyzed in the same way that 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 [17].

Source speaker's $\log F_0$ was converted by normalizing the $\log F_0$ 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 resynthesized using the source speaker's power and aperiodicity measures unmodified together with the converted spectral features and $\log F_0$.

For system K, 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 (Praat implementation: [20]).

3) Systems for Comparison: We intended to compare systems J and K with system F (described above), where transformation to the voice characteristics of the target speaker was performed using the same 94 min dataset but HMM-based adaptation rather than voice conversion methods. However, for system J the comparison is inexact, as system J does not impose any modification on source speaker's duration. Therefore, another system resembling system F in every way except for its duration model was constructed, system I. The only difference between systems F and I is that whereas in system F the average voice duration model is adapted to the duration characteristics of the child target speaker, in system I the same average voice duration model is adapted to the speaker characteristics of SLT.

C. Statistical Parametric Synthesis + Voice Conversion

Although systems J and K represent credible real-world configurations for voice-converted speech synthesis systems, comparison between them and systems F and I is compromised by two factors. First, as noted above, lower order spectral features were using in training the voice conversion components of J and K than were using in building voices F and I. Lower order features result in lower quality resynthesis which will adversely affect the performance of systems J and K in evaluation. Second, the fact that different systems were based on very different underlying voices also complicates the interpretation of evaluation results: J and K were based on a unit selection voice made from the data of a single speaker, whereas systems F and I were based on statistical parametric average voice made from the data of six speakers. Both these complicating factors mean that it would be hard to assess the relative performance gains or losses due to the use of either voice conversion or HMM adaptation techniques in isolation. Systems L-S were built in order to rectify this situation.

All of systems L-S were transformed from the same underlying voice, an HMM-based speaker dependent voice trained on the data of SLT following the procedure outlined for speaker dependent voices A, C, and E above, but using lower dimensional vectors of mel cepstral coefficients (25 static). The use of the same dimension of feature vector across both HMM-adapted and voice converted voices was intended to avoid the kind of bias towards system F in a comparison of systems F and K, inevitable due to the higher bit-rate vocoding used for F. The use of the same technology (statistical parametric) and training data (SLT) for the base voice was intended to remove the kind of discrepancies we would expect due to these factors not being kept constant in a comparison of systems F and K.

Furthermore, we wished to evaluate the contribution to system performance of each voice conversion component (GMM, scaling of F_0, stretching of duration) individually. Systems M, N, and O (see Table III) were all designed to be compared with system L—in each of these comparisons, the transformation method of a single voice component (spectrum, F_0, duration) is switched in isolation from HMM adaptation (with CSMAPLR and MAP as above) to the corresponding voice conversion method. The same scheme was applied in systems P–S, but whereas in systems L–O the 15-min target speaker dataset was used for transformation, in systems P–S the whole target speaker set of 94 min was used.

The procedures followed for training voice conversion functions of spectrum, F_0 and duration and applying them to test utterances used for systems M–O and Q–S were almost identical to those used for voice K, described above. The only difference was that, both for training and conversion, the spectral and F_0 features output by the base voice were fed directly into the voice conversion components: these features did not need to be extracted from waveforms.

The "pure HTS" voices L and P were built following the same general procedure outlined for speaker adaptive voices B, D, and F above. The only differences were that lower dimension feature vectors were used, and instead of an average voice, the base voice for the adaptation was a speaker-dependent voice (as already mentioned).

For all of voices L-S, aperiodicity measures were kept constant: in all cases they were generated by a model that had been adapted to the target speaker with HMM adaptation.

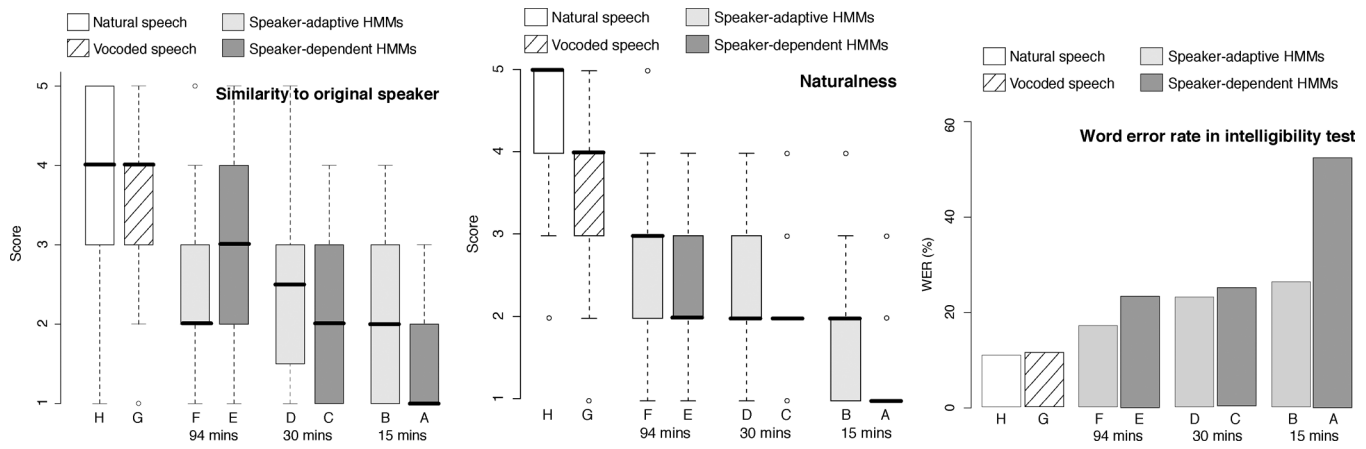

Fig. 3. Listening test results. Boxplot format follows [22]: "the median is represented by a solid bar across a box showing the quartiles; whiskers extend to 1.5 times the inter-quartile range and outliers beyond this are represented as circles."

TABLE IV
SENTENCES USED TO EVALUATE INTELLIGIBILITY OF NATURAL AND SYNTHETIC SPEECH

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.

IV. EVALUATION

A. Evaluation of Speaker-Dependent and Speaker-Adaptive Systems (A–F)

1) Procedure: The evaluation of the various systems was carried out using a similar protocol to the Blizzard Challenge [21]. Included in the set of "participants" (i.e., systems) were two benchmarks—natural speech and vocoded natural speech—where sentences held out of the corpus were used instead of synthesis. 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 F_0 value and higher formant frequencies of child speech, compared to adult speech, may cause spectral envelope estimation to be less accurate. The listening test, which was conducted via a web browser under quiet laboratory conditions using headphones, consisted of three sections. A Latin Squares design was employed meaning that in any given section, a single listener group heard every system once, each time with a different utterance. Every system was used to synthesize every utterance once within each section. We used a total of 48 paid listeners, all native speakers of English between the ages of 18 and 25.

Listeners were asked in Section I to rate the similarity of each stimulus to the original speaker. Two natural reference utterances were provided, which listeners could play at any time. Listeners could also hear 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." Second 2 followed the same format as the first, but this time listeners were asked to rate the naturalness of each stimulus on a five-point scale, with end points described to the listeners as "1—Completely Unnatural" and "5—Completely Natural." In Section III, 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 IV.

2) Results: The listening test data were analyzed using the same statistical techniques used in the Blizzard Challenge 2007 [22], and we present results in Fig. 3. Significant differences between systems are presented in Fig. 4. 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. WER was computed from a set of sentences of differing lengths, necessitated by the fact that these 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 normalized 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 Fig. 3 which receive partial support from the significance test. In most cases increasing the amount of training or adaptation data gives a higher median score in Sections I and II and a lower mean WER in Section III between systems of the same type, as we would expect.

In most cases, a speaker-adaptive voice yields higher median opinion scores and lower mean WER than a speaker-dependent

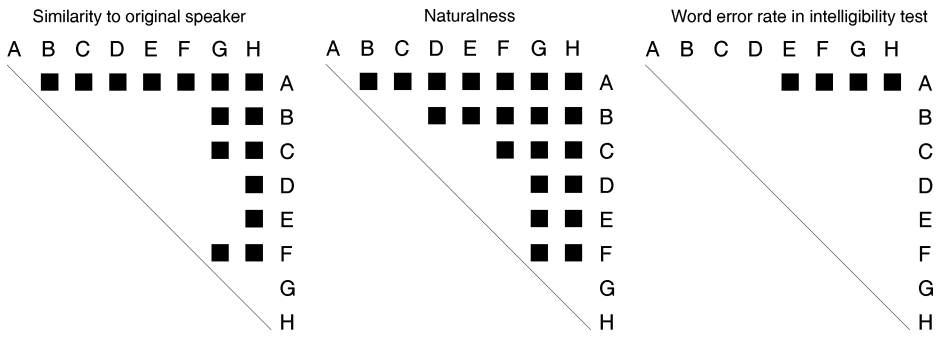

Fig. 4. Results of pairwise Wilcoxon signed rank tests between systems; a black square shows a significant difference between systems with $\alpha = 0.01$ (with Bonferroni correction).

voice trained on the same amount of data. 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. The phenomenon is observed in [12] and [23], and attributed to the fact that the average voice can be built from data covering a greater variety of contexts than is available for any single speaker. 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 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 I 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 whereas the median for the average voice-based system decreases slightly. 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.

We note that previous work has indicated that 100 utterances (approximately 6 min) of adaptation data are enough to adapt an average voice to the characteristics of a target speaker [23]. Fig. 3, however, shows that 15 min of child data achieve a median opinion score of only 2. We attribute this low score to the noisiness of the adaptation data and to the fact that the transformation attempted was unusually extreme: in [23], an average adult voice was adapted to another adult speaker, and not to a child.

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. These scores suggest that vocoding alone is causing degradation of the speech signal. Whether this degradation in quality (greater than we expected) 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. 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.

B. Evaluation of Unit Selection + Voice Conversion (J and K)

1) 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 analyzed and resynthesized as described in Section III-A1 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 could listen to the samples as much as they wanted. The ten "Goldilocks" sentences were synthesized 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 ten 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.

2) Results: Fig. 5 shows the results of the evaluation. Significant preferences were detected for all pairs of systems except I versus K.

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. However, interpretation of these preferences for HMM adapted over voice converted systems is complicated by the factors outlined in Section III above.

Both between HMM adapted systems F and I and between voice converted systems J and K we found significant preference for the system where duration transformation to the target speaker was performed (systems F and K, respectively). We note that the addition of duration transformation to the HMM adaptated system (F versus I) leads to a more extreme preference than in the case where uniform duration stretching is added to other voice converted system components (J versus K). In previous evaluations of HMM adapted voices [23], the inclusion of adaptation for duration leads to improved performance in subjective evaluation, but the preference in the present work for system F

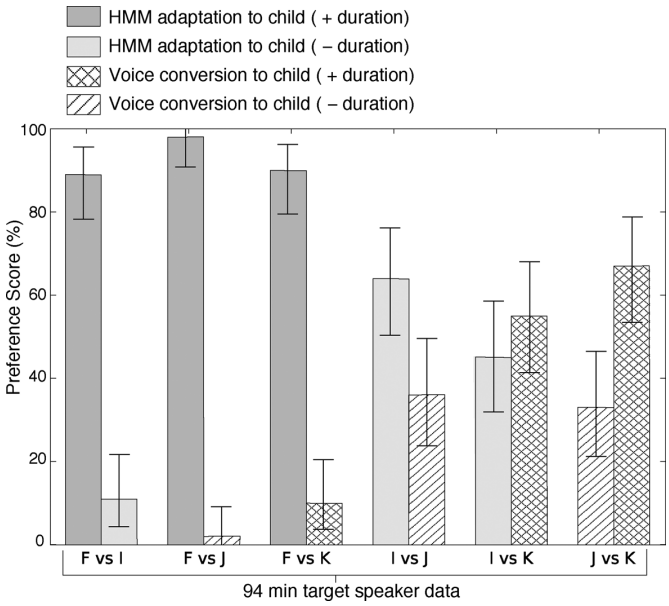

Fig. 5. Results of XAB test for speaker individuality, comparisons among systems F, I, J, and K. Vertical lines show 95% confidence intervals (with Bonferroni correction).

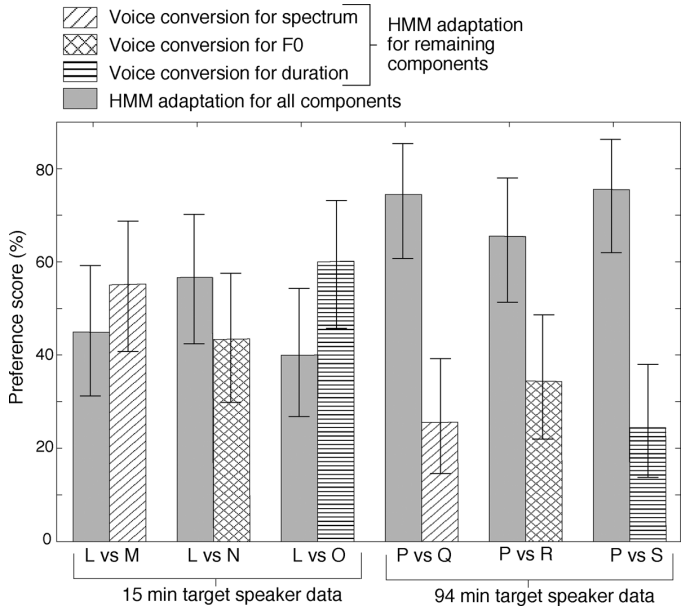

Fig. 6. Results of XAB test for speaker individuality; comparisons among systems L–S, Vertical lines show 95% confidence intervals (with Bonferroni correction).

over I is greater than that previous work led us to expect. We attribute the strength of this preference to the fact that speech duration patterns are very important in characterizing the child voice; these patterns distinguish a child's voice from an adult's to a much greater extent than they distinguish the speech of an adult from that of another adult speaker.

C. Evaluation of Individual Voice Conversion Components (systems L-S)

1) Procedure: Evaluation procedure was the same as that outlined in Section IV-B; here also, six pairwise comparisons were made, but this time the comparisons each focused on the impact of a single voice conversion component on system performance. Two slight divergences from the procedure in Section IV-B were made: the natural reference samples were original waveforms, not vocoded speech, and nine listeners performed the evaluation, not ten.

2) Results: Fig. 6 shows the results of the evaluation. In all the voices transformed with the small (15 min) dataset (L–O), no significant difference is detected between transformation components based on HMM adaptation and those based on voice conversion methods (GMM and uniform scaling of $\log F_0$ and duration). However, when 94 min of data is available (P–S), there is a significant preference for the HMM adaptation technique in every case. These findings are consistent with our expectations: given sufficient data we would expect the HMM adaptation techniques to give better results due to the fact that the decision tree used incorporates high-level linguistic and prosodic information. It is also our experience that the performance of shallow voice conversion methods—informed by acoustic features only—degrades slowly as the amount of training data available becomes very small. Voice conversion systems can be trained using the techniques tested here with as few as 30 target speaker utterances and still perform respectably. An interesting topic for future work would be to make similar comparisons between HMM adaptation and voice conversion methods as those outlined here, but with target speaker datasets smaller than 15 min.

V. Conclusion

This paper described the application of existing HMM-based speech systems to the synthesis of a child's speech. Both speaker-dependent voices and speaker-adapted voices were built. Additionally, we compared the performance of systems transformed to the child target speaker with HMM adaptation techniques to that of those where one or more components of the voice are transformed using techniques from voice conversion.

Although the child speech data has poor coverage of the phonetic/prosodic units of the language, an inconsistent reading style, and imperfect recording conditions, it is feasible to build child voices on the database by using the robust framework of HMM-based speech synthesis. In the evaluation, it was found that adult average voices adapted to the target child speaker data generally outperformed child speaker-dependent voices. This had been found to be true in the case of adult target speakers in previous work. However, we found that more target speaker data were needed to get reasonable speaker similarity rating when adapting to the child target speaker than in the previous work with adult target speakers. Also, speaker similarity ratings level off with the largest size of target speaker database, and better ratings are obtained for the speaker-dependent system. This implies that better average voice models for child speech are required.

Our comparison of the effectiveness of HMM adaptation techniques with voice conversion style techniques for imposing target speaker characteristics on a base voice sheds light on

their performance differences. When the adaptation data is restricted to 15 min, there was no significant preference for either HMM adaptation or voice conversion methods. This is also underpinned by the theoretical background that both the techniques use similar feature linear transforms when available data is limited. On the other hand, more importantly, HMM adaptation was preferred in every case when using the full target speaker corpus. This is because relatively large amounts of data enable extensive use of the decision tree that incorporates high-level linguistic and prosodic information in speaker adaptation. Furthermore the adaptation of durational characteristics was found to have a greater impact on listener preference than we were led to expect from previous work on adaptation to adult speakers.

Our future work is to build child average voice models using Speecon databases [24] where 60 children (8 to 15 years old) read 30 phonetically rich sentences. Although the databases do not cover seven-year old children, average voice models trained on the databases would provide better prior information than adult average voice models used for our experiments.

This work has evaluated the performance on child speech of techniques that—while not adult-specific—have nevertheless been developed and tuned principally on adult speech. Another possible direction for future work is to try child-specific speaker adaptation algorithms, such as the one proposed in [25].

Audio examples of the child synthetic speech built are available at http://homepages.inf.ed.ac.uk/s0676515/child_speech

ACKNOWLEDGMENT

The authors would like to thank R. Zundel, daughter of K. Berkling, for her contribution to this research through many hours of reading and recording.

REFERENCES

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

[2] O. Watts, J. Yamagishi, S. King, and K. Berkling, "HMM adaptation and voice conversion for the synthesis of child speech: A comparison," in *Proc. Interspeech'09*, Brighton, U.K., Sep. 2009, pp. 2627–2630.

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

[4] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[5] J. Yamagishi, Z. Ling, and S. King, "Robustness of HMM-based speech synthesis," in *Proc. Interspeech'08*, Brisbane, Australia, Sep. 2008, pp. 581–584.

[6] 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*, Brisbane, Australia, Sep. 2008.

[7] J. Yamagishi, M. Tachibana, T. Masuko, and T. Kobayashi, "Speaking style adaptation using context clustering decision tree for HMM-based speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '04).* , May 2004, vol. 1, pp. 1–5.

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

[9] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *J. Acoust. Soc. Amer.*, vol. 105, no. 3, pp. 1455–1468, 1999.

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

[11] 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.*, vol. E90-D, no. 1, pp. 325–333, Jan. 2007.

[12] 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.

[13] 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 Commun.*, vol. 27, pp. 187–207, 1999.

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

[15] 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. Audio, Speech, Lang. Process.*, vol. 17, no. 1, pp. 66–83, Jan. 2009.

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

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

[18] A. R. Toth and A. W. Black, "Incorporating durational modification in voice transformation," in *Proc. Interspeech'08*, Sep. 2008.

[19] A. W. Black and K. A. Lenzo, "Building Synthetic Voices." 2007 [Online]. Available: http://festvox.org/bsv/

[20] P. Boersma and D. Weenink, "Praat: Doing Phonetics by Computer." 2005 [Online]. Available: http://www.praat.org/

[21] M. Fraser and S. King, "The Blizzard challenge 2007," in *Proc. BLZ3–2007 (in Proc. SSW6)*, Aug. 2007.

[22] 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.

[23] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Trans. Inf. Syst.* , vol. E90-D, no. 2, pp. 533–543, Feb. 2007.

[24] D. Iskra, B. Grosskopf, K. Marasek, H. Heuvel, F. Diehl, and A. Kiessling, "Speecon – speech databases for consumer devices: Database specification and validation," in *Proc. LREC'02*, Las Palmas, Spain, 2002, pp. 329–333.

[25] S. Wang, Y.-H. Lee, and A. Alwan, "Bark-shift based nonlinear speaker normalization using the second subglottal resonance," in *Proc. Interspeech'09*, Brighton, U.K., Sep. 2009, pp. 1619–1622.

Oliver Watts received the M.Sc. degree in speech and language processing from the University of Edinburgh, Edinburgh, U.K., in 2007. He is currently pursuing the Ph.D. degree at the Centre for Speech Technology Research, University of Edinburgh, working on speech synthesis in languages where few resources are available.

Junichi Yamagishi (M'05) received the B.E. degree in computer science and the M.E. and Dr.Eng. degrees in information processing from the Tokyo Institute of Technology, Tokyo, Japan, in 2002, 2003, and 2006, respectively.

He was an Intern Researcher at ATR Spoken Language Communication Research Laboratories (ATR-SLC) from 2003 to 2006. He was a Visiting Researcher at the Centre for Speech Technology Research (CSTR), University of Edinburgh, Edinburgh, U.K., from 2006 to 2007. He is currently a Senior Research Fellow at the CSTR and continues the research on the speaker adaptation for HMM-based speech synthesis in an EC FP7 collaborative project called the *EMIME* project (www.emime.org). His research interests include speech synthesis, speech analysis, and speech recognition.

Dr. Yamagishi received the Tejima Doctoral Dissertation Award 2007 with his doctoral dissertation *Average-voice-based speech synthesis* which pioneered the use of speaker adaptation techniques in HMM-based speech synthesis. He held a research fellowship from the Japan Society for the Promotion of Science (JSPS) from 2004 to 2007.

He is a member of ISCA, IEICE, and ASJ.

Simon King (M'95–SM'08) received the M.A.(Cantab) degree in engineering and the M.Phil. degree in computer speech and language processing from the University of Cambridge, Cambridge, U.K., in 1992 and 1993, respectively, and the Ph.D. degree in speech recognition from the University of Edinburgh in 1998.

He has been involved in speech technology since 1992, and has been with the Centre for Speech Technology Research, University of Edinburgh, since 1993. He is a Reader in Linguistics and English Language and an EPSRC Advanced Research Fellow. His interests include concatenative and HMM-based speech synthesis, speech recognition, and signal processing, with a focus on using speech production knowledge to solve speech processing problems.

Dr. King is a member of ISCA, serves on the steering committee for SynSIG (the special interest group on speech synthesis) and co-organizes the Blizzard Challenge. He is an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING.

Kay Berkling (M'05–SM'06) received the B.S. degree in mathematics and computer engineering and the B.A. degree in German and French from Syracuse University, Syracuse, NY, in 1991, and the Ph.D. degree in automatic language identification from The Center for Spoken Language Understanding, Oregon Institute of Science and Technology, Portland, in computer science and engineering.

She completed three postdoctoral positions, where her research concentrated on foreign accent identification in English speech. Following an appointment in the Speech Group at the MIT Lincoln Laboratory, Lexington, MA, she joined the UBS Innovation Lab, Ubilab, in 2000. After working in the financial industry of Switzerland as an e-business consultant, she took a position at Polytechnic University of Puerto Rico, San Juan, as a Professor in 2004. She lectures in the Department of Computer Science and has recently hosted ASRU 2005 and Odyssey 2006. Research interests include computer-aided language teaching and incorporating linguistic knowledge into speech recognition systems as well as accent-, dialect-, and language-identification systems. Her goal is to bridge the educational gap of non-native speakers of any language through the use of tools employing speech recognition.