Robust recognition of children's speech

Alexandros Potamianos[1], *Member IEEE*, and Shrikanth Narayanan[2], *Senior Member IEEE*
[1]Dept. of Electronics & Computer Engineering, Technical Univ. of Crete, Chania 73100, Greece
[2]Dept. of Electrical Engineering, Univ. Southern California, Los Angeles, CA 90089, U.S.A.
potam@telecom.tuc.gr, shri@sipi.usc.edu

Abstract

Developmental changes in speech production introduce age-dependent spectral and temporal variability in the speech signal produced by children. Such variabilities pose challenges for robust automatic recognition of children's speech. Through an analysis of age-related acoustic characteristics of children's speech in the context of automatic speech recognition (ASR), effects such as frequency scaling of spectral envelope parameters are demonstrated. Recognition experiments using acoustic models trained from adult speech and tested against speech from children of various ages clearly show performance degradation with decreasing age. On average, the word error rates are two to five times worse for children speech than for adult speech. Various techniques for improving ASR performance on children's speech are reported. A speaker normalization algorithm that combines frequency warping and model transformation is shown to reduce acoustic variability and significantly improve ASR performance for children speakers (by 25–45% under various model training and testing conditions). The use of age-dependent acoustic models further reduces word error rate by 10%. The potential of using piece-wise linear and phoneme-dependent frequency warping algorithms for reducing the variability in the acoustic feature space of children is also investigated.

Keywords

Automatic Speech Recognition, Children, Formant scaling, Robustness, Speaker normalization, Vocal tract normalization.

EDICS: 1-RECO, 1-ANLS

I. Introduction

Automatic speech recognition (ASR) for children speakers is a challenging problem with many potential applications in education, entertainment and communication services. A significant amount of literature exists on acoustic and linguistic analyses of both normal and pathological speech production of children, including comparisons with adult speech patterns [6], [8]. Such studies have investigated changes in the spectral and temporal characteristics of the speech signal across various age groups. There has been, however, relatively little published work on issues and algorithms related to automatic recognition of children's speech [12], [15].

Most speech recognition systems that have been deployed target adult users and experience severe ASR performance degradation when used by children users. For example, while analyzing ASR performance of the live usage data obtained from the MIT Jupiter system, Zue et al [20, Fig. 9] found that the in-vocabulary word error rate for children was almost twice that for adult users. There have been a few ASR application prototypes that have specifically aimed at children such as word games for pre-schoolers [18], aids for reading [10] and pronunciation tutoring [16]. In all these applications, the use of speech recognition has been constrained – either in terms of highly limited vocabulary (e.g., just "yes" or "no" in games) or dependence on perfect knowledge of the lexical items during recognition such as in reading applications (thereby simplifying the problem of

recognition to that of forced alignment). In this paper, the problem of robust speaker-independent recognition of children's speech is addressed.

In a study of ASR in elderly and children talkers [19], ASR performance was shown to degrade when children's speech was tested against models derived from adult speech for a connected digits task. Similar results were later reported by [1], [15], [2]. A simple transformation procedure from the children to the adult acoustic feature space was implemented in [1] to compensate for such mismatches. A detailed investigation of the age-dependent effects in ASR performance under a variety of model conditions and speaker normalization was reported in [15]. This paper expands upon the results of [15] providing details on issues related to ASR of children speech and techniques toward addressing some of those issues.

The rest of the paper is organized as follows. In Sec. II, an analysis of the speech signal characteristics in the context of ASR is provided. Specifically, a detailed investigation of frequency scaling effects in the speech spectrum with age is reported. Acoustic modeling for ASR is considered in Sec. III. Databases used are summarized in Sec. III-A while baseline speech recognition results highlight ASR performance as a function of the speaker's age are presented in Sec. III-B. The effects of frequency warping for speaker normalization are analyzed in Sec. IV-A. Algorithms for combined frequency warping, model transformation and model selection in a maximum likelihood framework are given in Sec. IV-B. Experimental results for connected digits and command/control tasks are provided in Sec. V followed by conclusions in Sec. VI.

II. Analysis of children's speech

The spectral and temporal characteristics of children's speech are highly influenced by growth and other developmental changes and are different from those of adult speakers. These differences are attributed mainly to anatomical and morphological differences in the vocal-tract geometry, less precise control of the articulators and a less refined ability to control suprasegmental aspects such as prosody. In a key study by Eguchi and Hirsh [3], and later summarized by Kent [6], age-dependent changes in formant and fundamental frequency measurements of children speakers ages three to thirteen were reported. Important differences in the spectral characteristics of children voices when compared to those of adults include higher fundamental and formant frequencies, and greater spectral *variability* [3], [6], [8]. Parametric models for transforming vowel formant frequency of children speakers to the adult speaker space (vowel formant frequency normalization) have also been considered, for example in [4], [8]. Similarly, a detailed comparison of temporal features and speech segment durations for children and adult speakers can be found in [6], [17], [8]. Again, distinct age-related differences were found: On average, the speaking rate of children is slower than that of adults. Further, children speakers display higher variability in speaking rate, vocal effort, and degree of spontaneity.

Most of the early acoustic studies were somewhat limited in terms of the size of the databases analyzed, especially in terms of the number of subjects. Furthermore, enabling ASR was not among the major goals of these early studies. In a related study by the authors, variations in the temporal and spectral parameters of children's speech were investigated using a comprehensive speech data corpus (23454 utterances) obtained from 436 children ages between 5 and 18 years and 56 adults [8]. In the next section, key findings from that study that are relevant to automatic speech recognition, including results on formant scaling, are summarized.

A. Age-dependencies in speech acoustic characteristics

To obtain insights into age-dependent behavior in the magnitude and variance of the acoustic parameters, measurements of spectral and temporal parameters were made through a detailed

analysis of the American English vowels [8]. Results showed a systematic decrease in the values of the mean and variance of the acoustic correlates such as formants, pitch and duration with age, with their values reaching adult ranges around 13 or 14 years. A specific result that is relevant for ASR is the scaling behavior of formant frequencies with respect to age. As can be seen from Fig. 1(a), the vowel space (quadrilateral boundaries marked by the four-point vowels /AA, IY, UW, AW/ in the F2-F1 plane plotted in mel frequency scale) changes with increasing age in an almost linear fashion. The movement of the vowel quadrilateral is in the direction toward smaller F2-F1 values with increasing age corresponding to the lengthening of the vocal tract associated with growth. Also, it can be noticed that the the vowel space becomes more compact with increasing age indicating a decreasing trend in the dynamic range of the formant values. A more detailed account of the scaling behavior can be obtained by plotting the variation in the formant scaling factors (calculated as a ratio of average formant frequency values for a specific age group to the corresponding values for adult males). The plots in Fig. 1(b) show a distinct and an almost linear scaling of each of the first three formant frequencies with age. The scaling trend for females and males diverges significantly after puberty suggesting underlying differences in their anatomical growth patterns. Moreover, the first three formants scale more uniformly for males. Formant frequencies of females, on the other hand, show a more nonlinear scaling trend for the various formants especially after puberty.

The intra-speaker variability (i.e., within subjects) was larger for young children, especially for those under 10 years. Fig. 2 shows a decreasing trend in intra-subject variability with age in terms of cepstral distance measures of variability both within a token and across two repetitions. It is generally believed that both the acoustics and linguistic correlates of children speech are more variable than those of adults. For example, the area of the F1-F2 formant ellipses was larger for children than for adults for most vowels [3]. Children speech also contains more disfluencies and extraneous speech especially in human-machine interactions [18]; such results are however highly dependent on how the data was collected (read speech vs spontaneous speech). Some insights regarding the acoustic characteristics of children's spontaneous speech can be obtained from the results of another related study [11]. The analysis is based on data obtained from a Wizard of Oz spoken dialog experiment using 160 children playing a voice-activated computer computer game (a total of 22000 utterances from 7-14 year olds were used for the following analysis). The average sentence duration was about 10% longer for younger children. As a result, the speaking rate for the 11-14 year-olds was about 10% higher than for the younger group which is in agreement with the results on read speech [8]. An important aspect of spontaneous speech is the prevalence of disfluencies. Disfluencies and hesitations in the speech data were analyzed as a function of age and gender. Mispronounciations, false-starts, (excessive) breath noise and filled pauses (e.g., um, uh) were manually labeled for a subset of the data (22422 utterances). About 2% of the labeled utterances contained false-starts and 2% contained (obvious) mispronounciations. Breathing and filled pauses were found in 4% and 8% of the utterances, respectively. While no gender dependency was found for any of the disfluency measures, there was a distinct age dependency. The frequency of mispronounciations was almost twice as high for the younger (8-10 years) age group than for the older group (11-14 years). Breathing noises occurred 60% more often for younger children. Surprisingly, this trend was reversed for filled pauses which occurred almost twice as often for the 11-14 age group.

B. Implications to ASR

There are several implications that the acoustic characteristics of children's speech discussed above have on automatic speech recognition. First, the age-dependent scaling in formant frequencies introduces variability in the spectral features across age groups. As a result, if a model for ASR

is based on data pertaining to a certain age-group and tested against data belonging to other age groups, the mismatch between the model and data results in performance degradation. If the scaling factor relation between the reference data and new data were known then, in principle, one could re-scale the spectral features to reduce any mismatch. Various issues in implementing frequency warping for model normalization are investigated in Sec. IV-A. In ASR, scaling factors between reference and test data are often not known a priori and have to be estimated from data during recognition. In this work, a maximum likelihood approach for parametric speaker normalization is adopted as described in Sec. IV-B.

A second major challenge in acoustic modeling for ASR is the spectral and temporal variability in children's speech. Increased variability in formant values results in greater overlap among phonemic classes for children than for adult speakers, and makes the classification problem inherently more difficult. Further, the range of values for most acoustic parameters is much larger for children than for adults. For example, five-year old children have formant values up to 50% higher than male adults [8]. The difficulty for spectral-feature based pattern classification due to increased dynamic range of acoustic parameters is illustrated in the F1-F2 formant space shown in Fig. 3 for various vowels spoken by adult and children. The sizes of the phonemic classes (represented by the area of the ellipses in the F1-F2 plot) for children speakers is much larger than for adults, which results in significant overlap among classes. The combination of a large acoustic parameter range and increased acoustic variability can seriously degrade ASR performance. In Sec. IV, speaker normalization procedures and age-dependent acoustic modeling are used to reduce variability and increase resolution between classes.

Third, there are certain fundamental limitations on feature extraction from the speech of young children. The main goal of the ASR feature extraction stage is to decompose the speaker-dependent information (e.g., fundamental frequency F0) from the phoneme-dependent information (e.g., formant frequencies) and retain the latter. This task is more difficult for children voices because the fundamental frequency and the formant bandwidths are of comparable magnitude. As a result, speaker dependent information exists in the feature vectors derived from children speech which, in turn, results in degradation of the classification performance. Another factor is the effect of finite spectral bandwidth. For example, in telephone speech, a large spectral chunk containing the high-frequency formants is lost due to band-limiting. As a result, the acoustic information available in the telephone channel bandwidth is less for children than for adult speakers, e.g., typically only 2-3 formants exist in the 0.3-3.2 kHz range for children speech compared to 3-4 for adults. Thus, the sparse sampling of the spectrum (due to high F0 values) and relatively few formants in the given bandwidth (due to high formant values) in children's speech pose fundamental limitations on the amount of phoneme-dependent information available at the ASR front-end. In this work we do not explicitly investigate the effects of different bandwidths and signal parameterization on ASR. Instead, we treat these as yet another source of mismatch and attempt to address them through combined speaker normalization and adaptation during acoustic modeling for ASR (Sec. IV-B).

Finally, we consider issues in ASR for children that relate to the effects of spontaneity and linguistic variability in children's speech (e.g. disfluencies and extraneous "out of domain" speech). Although disfluencies and hesitation phenomena occur more frequently in children than in adults, our experiments showed that ASR performance does not suffer significantly due to these effects, hence requiring no special acoustic modeling strategies other than the commonly used garbage models in ASR. As for the effects of linguistic variability, they are better handled at the language modeling and dialog interaction levels, topics that are outside the scope of this paper. The main focus of the rest of the paper is on acoustic modeling and experiments related to robust ASR of children's speech.

III. Acoustic Modeling for ASR

As discussed in the previous section, the acoustic characteristics of children change rapidly as a function of age and are different from those of adults. Further, the intra-speaker acoustic variability is much higher for younger children than for teenagers and adults. Thus it is expected that automatic speech recognition performance to decrease as a function of age especially if the reference models are based on adult speakers. In this section, details of experiments to determine ASR performance as a function of speaker's age are described. For this purpose continuous hidden Markov models (HMMs) were trained using utterances from adult and children speakers with data collected over the public switched telephone network.

A. Databases

Several databases were collected to enable the experiments described in this paper. A summary of the databases used for training and testing the acoustic models, and for speech analysis purposes is provided in Table I. The databases allowed experimentation with both connected digit and subword-based speech recognition tasks. Prior to the data collection, children were provided with instructions through their parents, along with the speech material to be read (digit strings of length 1,3,4,7 or 11 digits, phonetically balanced sentences, and a list of short command and control words), for calling a toll-free number to do the recordings. A simple touch-tone interface was devised to automate the data collection. The data were manually verified for transcription accuracy.

In addition, the database used for speech analysis in Sec. II (referred to as MicI and MicII in Table I) was used to investigate issues in frequency warping since it provided a well balanced corpus of the phonemes in American English [9]. The database was collected from 436 children of ages five through eighteen, and 56 adults. The speech material consisted of ten monophthongs and five diphthongs of American English vowels (embedded in a carrier word) and five phonetically–balanced sentences, repeated twice by all subjects.

B. Baseline ASR Experiments

To illustrate age-dependent effects on ASR performance, the baseline experiments focussed on a connected digit recognition task with acoustic models built using adult speech data and tested

Name	Speaker Population	Content	No. of speakers	No. of strings
TRAINING				
DgtI	Adults	digits	3026	4781
DgtII	10-17 yrs.	digits	1234	5767
SubwI	Adults	phrases	242	12144
SubwII	10-17 yrs.	phrases	1234	14267
TESTING				
DgtIII	6-17 yrs.	digits	501	2656
CommI	6-17 yrs.	commands	501	3554
CommII	10-17 yrs.	commands	1234	7436
ANALYSIS				
MicI	5-18 yrs.	phonemes	436	13080
MicII	5-18 yrs.	sentences	436	4360

TABLE I

Training, testing and analysis databases.

against children's speech (as will be shown later in Fig. 9, results for subword recognition were similar). Context-dependent whole-word hidden Markov digit models were built using 3 states with 6 Gaussian mixtures/state.

Fig. 4 (a) and (b) show digit recognition rates for male and female speakers as a function of age for a connected digit recognition task based on corpus DgtIII. Separate HMMs – labeled "Adult HMM" and "Child. HMM" – were trained using utterances from adult corpus DgtI and children corpus DgtII. The results show, as expected, the error rates increase with decreasing age. The performance is especially poor for children younger than 10 years under both matched and (especially for) mismatched training and testing conditions: The error rate is approximately ten times higher for very young children than for adults. ASR performance reaches adult levels around thirteen or fourteen years of age, which is in agreement with the observation in [8] that by the age of fourteen both the mean and variance of most acoustic characteristics have reached levels similar to that of adult speech. Overall recognition performance for children speakers is about four times worse than for adults. For mismatched training and testing conditions ("Adult HMM") word error rate is about two to three times higher than for matched conditions ("Child. HMM"). We also observed that a small improvement of 5-10% was achieved by using context-dependent (vs. context-independent) model units, which is consistent with the observation in [8] that young children (ages 5-12) have not fully developed their co-articulation skills. In summary, the major reason for performance degradation for younger speakers is due to increased acoustic variability and the large range of acoustic parameters (as discussed in Sec. II). Next we present techniques aimed at reducing the mismatch and variability.

IV. Speaker Normalization and Model Adaptation

Most state-of-art ASR systems use front-end features extracted from the short-time average of the smooth spectral envelope such as LPCC or MFCC. In Sec. II, the age-dependent scaling effects in the speech spectral parameters were shown. For a given model condition (e.g., models trained from adult speech), such scaling effects imply spectral mismatches and in turn, degradation in ASR performance. In this section, we demonstrate that frequency warping can substantially decrease the average spectral difference between children and adult speech. First, we present a systematic analysis of frequency warping effects across speakers and phone types. The average scaling factors between children and adult speech are computed for all phonemes and it is shown that the inter-phoneme scaling factor variation is relatively small across vowels. We then present a maximum likelihood approach for combined frequency warping and spectral shaping.

A. Analysis of Frequency Warping in Children's Speech

The high quality microphone speech database MicI [9], [8] was used for providing insights into the effects of frequency warping. For this analysis, a Euclidean cepstrum distance was defined to measure the similarity between two speech frames. Specifically, to determine the distance between two speech segments A and B: (i) the logarithm of the spectral envelope was computed for each of the speech segments using a mel-spaced (24 filter) filterbank covering the frequency range from 200 Hz to 4 kHz (filters are 50(ii) the spectral envelope was normalized by subtracting the corresponding average spectral log energy (zero mean), (iii) the inverse cosine transform of the spectral envelope was computed and (iv) the Euclidean distance between the vectors of the first twelve inverse cosine (cepstrum) coefficients was calculated. The distance measure D is defined as

$$D = \sum_{n=1}^{12} (c_n^A - c_n^B)^2 \tag{1}$$

where c_n^A is the nth inverse cosine transform coefficient (or the nth cepstrum coefficient) for speech segment A. The distance measure D is proportional to the logarithm of the probability scores of an HMM-based classifier used for the recognition experiments in Sec. V, provided that the variances of all features and for all classes are assumed to be equal. The dynamic features (first and second cepstrum time discrete derivatives) are not incorporated in the distance metric since only the steady-state portion of the phonemes is considered. It should be pointed out that a measure such as Mahalanobis distance would provide a more accurate evaluation in terms of accounting for the different variances of different features. Nevertheless, the computationally simpler approach was found to be sufficient to provide the necessary insights on frequency warping.

A.1 Linear Frequency Warping: Vowels

We first investigate the effects of frequency warping at the phoneme level. We demonstrate a substantial reduction in the Euclidean cepstrum distance between the vowel segments of children and adult speech before and after linear frequency warping. Frequency warping is performed as follows: (i) for each monophthongal vowel and for each age and gender group, the average spectral envelope is computed, (ii) the optimal scaling factor is computed (for each vowel, speaker's age, and gender) so that the Euclidean cepstrum distance between the warped children spectral envelope and the corresponding adult reference spectral envelope is minimized; optimization is achieved by searching exhaustively in the interval of warping factors ranging from 0.7 to 1.15, where 1 corresponds to no warping, and (iii) the average spectral envelope for each vowel, speaker's age, and gender is warped according to the optimal warping factor. Frequency warping is implemented by re-sampling the spectral envelope at linearly scaled frequency indices.

In Fig. 5(a)-(d) the Euclidean cepstrum distance between male children and male adult speakers is shown before and after frequency warping averaged over children ages (a) 5-7 years (b) 8-10 years (c) 11-13 years and (d) 14-16 years. The distance between the average vowel spectral envelopes for children and adult speakers decreases approximately ten times when frequency warping is applied. The average distance (before warping) between children and adult speakers decreases rapidly as a function of age and becomes negligible for children over 14 years of age. The "after warping" distance follows a similar decreasing trend with increasing children's age which suggests that linear frequency warping is useful but only a first step towards reducing the acoustic mismatch between children and adult speakers. In Fig. 6(a), (b) the average and standard deviation of the percent reduction in spectral distance between utterances of male children and adult speakers due to warping is shown for two age groups. In these plots the scaling factor and distance reduction *is computed for each phonemic instance* (as opposed to Fig. 5 where distances are computed per phoneme). Note that linear frequency warping is a more effective normalization procedure for front vowels rather than back vowels for both age groups. Further, for the younger children the percent distance reduction due to warping is higher and more consistent (smaller standard deviation) among different speakers.

In Fig. 7, the warping factors obtained for male and female children, relative to adult male and female speakers, are shown as a function of age. The optimal warping factors are computed for each vowel, age and gender group as described in the previous paragraph and then averaged over the ten monophthongal vowels of American English. The resulting warp factors are directly comparable with the formant scaling results computed for this database in [8]. The warping factor contours as a function of age are very similar in (a) and (b) independent of the reference spectrum (male or female adult speakers). It can be inferred from Fig. 7 that the spectral characteristics reach adult levels around age 14 for females and around age 15 for males. Further, the growth spurt

around 11 to 13 years of age is clearly shown for male children[1]. It is interesting to note that the warping factors for male children ages 5-9 are consistently larger than those of female children of the same age. In [8], there was a significant difference between the first formant frequency scaling factors and absolute first formant values between male and females below 10 years of age, while F2 and F3 values were very similar. There is no clear physiological explanation for the difference in spectral scaling factors and F1 values for young male and female children. The difference could be attributed to speech style differences, e.g., male and female children trying to imitate adult male and female speaker characteristics, respectively.

A.2 Phoneme-Dependent Frequency Warping

In the previous experiment, a global warping factor was used i.e., an averaged value for the warping function was applied to all phonemes. Here we investigate the validity of this assumption by comparing warping factors and spectral distances before and after frequency warping for different phonemic groups. In Fig. 8(a), the optimal warping factors for male children of various age groups relative to adult male speakers are shown for vowels (monophthongs and diphthongs), nasals, glides and fricatives. The warping factors were computed as described in the previous sections and averaged over age groups 5-8, 9-12 and 13-16 years. The inter-vowel warp factor variability for each of the age groups is relatively small and is greatest for the 5-8 age group. Warping factors typically are smaller for diphthongs, glides and nasals than for vowels; the values for fricatives were the smallest amongst all the phonemic classes as expected. All phonemes displayed similar patterns in the variation of warping factors as a function of age. This suggests that that the same type of warping function can be used for all phonemes as a good first order approximation.

In Fig. 8(b) the average Euclidean cepstrum distance (similar to the log likelihood used for ASR) between male children ages 5–8 and adults, before and after frequency warping is computed for each phoneme. The simple linear frequency warping (by the amount in Fig. 8(a)) is shown to be very efficient in reducing acoustic mismatch between the young children and adult speakers for most phonemic classes. Note the relatively relative large distance reduction for vowels and glides, a small distance reduction for nasals, and practically no distance reduction for fricatives (with the exception of /sh/). In summary, the use of phoneme-dependent warping factors when applying frequency normalization can further reduce spectral mismatch between the original and target group of speakers than using a global value for the warping factor. In practice, such gains can be rather limited because of the great number of scaling factors that have to be estimated. Alternatively one can compute class-dependent warping factors using three or four broad phonemic classes as suggested by the patterns in Fig. 8(a).

B. Algorithms for Speaker Normalization

In this section, speech recognition performance is improved by reducing the mismatch between the acoustic models and test utterances, and by reducing the inherent acoustic variability of the models. As indicated by the analysis in the previous section, frequency warping helps significantly to reduce spectral differences caused by age-dependent effects. The frequency warping approach to speaker normalization compensates mainly for inter-speaker vocal tract length variability by linear warping of the frequency axis by a factor α ($\alpha = 1$ corresponds to no warping). We adopt the approach proposed in [7] for warping an utterance according to a parametric transformation $g_\alpha()$ in order to maximize the likelihood of the observation with respect to a model. Frequency

[1]The low value of the warping factor for male children of age 14 is not an artifact of the warping factor computation process. It is supported by formant measurements on this database.

warping is implemented in the mel-frequency filterbank front-end by linear scaling of the spacing and bandwidth of the filters. Scaling the front-end filterbank is equivalent to re-sampling the spectral envelope using a compressed or expanded frequency range. The speaker normalization algorithm works as follows. For each utterance, the optimal warping factor $\hat{\alpha}$ is selected from a discrete ensemble of possible values so that the likelihood of the warped utterance is maximized with respect to a given HMM and a given transcription. The values of the warping factors in the ensemble typically vary over a range corresponding to frequency compression or expansion of approximately ten percent. The size of the ensemble is typically ten to fifteen discrete values. Let $X^\alpha = g_\alpha(X)$ denote the sequence of cepstrum observation vectors where each observation vector is warped by the function $g_\alpha()$, and the warping is assumed to be linear. If λ denotes the parameters of the HMM model, then the optimal warping factor is defined as

$$\hat{\alpha} = \arg\max_{\alpha} P(X^\alpha|\alpha, \lambda, H) \tag{2}$$

where H is a decoded string obtained from an initial recognition pass. The frequency warped observation vector X^α is used in a second recognition pass to obtain the final recognized string. Note that the procedure is computationally efficient since maximizing the likelihood in Eq. 2 involves only the probabilistic alignment of the warped observation vectors X^α to a single string H.

B.1 Combining Frequency Warping and Spectral Shaping for Speaker Normalization

This section describes a simple method for implementing a parametric linear transformation on the HMM model in conjunction with a parametric frequency warping of the input utterance in a single statistical framework. The method can be interpreted as a means for expanding the ensemble of alternatives that are being evaluated during adaptation thus obtaining a better match between the input utterance and the model. Speech collected from children exhibits consistent spectral differences (other than formant scaling), when compared to adults, e.g., spectral tilt. Spectral shaping can be used to compensate for such spectral trends thus reducing the mismatch and the inherent variability in children acoustic models. Spectral shaping can be used in conjunction with frequency warping to achieve incremental reduction of mismatch and model variability.

There is a large class of maximum likelihood based model adaptation procedures that can be described as parametric transformations of the HMM model parameters. Let $\lambda_\gamma = h_\gamma(\lambda)$ denote the model obtained by a parametric linear transformation $h_\gamma()$. The form of the transformation depends on a number of issues including both the nature of the sources of variability and the amount of data available for estimating the parameters of the transformation. However, the same maximum likelihood criterion can be used for estimating γ as was used for estimating α:

$$\hat{\gamma} = \arg\max_{\gamma} P(X|\gamma, \lambda_\gamma, H) \ . \tag{3}$$

Our goal is to combine the frequency warping and model adaptation methods in a maximum likelihood framework. The optimal parameters of the model transformation $\hat{\gamma}$ and the frequency warping $\hat{\alpha}$ can be simultaneously estimated so that

$$\{\hat{\alpha}, \hat{\gamma}\} = \arg\max_{\{\alpha,\gamma\}} P(X^\alpha|\alpha, \gamma, \lambda_\gamma, H) \ . \tag{4}$$

A computationally efficient implementation of the combined procedure can be used assuming a simple definition for the model transformation, $h_\gamma()$. Specifically, if the model transformation corresponds to a single fixed transformation applied to all HMM means, it can be applied to the

observation sequence instead of the HMM.[2] Similarly, instead of building "warp class" models, frequency warping can be applied directly on the observation sequence during testing. This simplifies both the computational load and the memory requirements of the speaker normalization and adaptation procedure. As before, we attempt to simultaneously optimize the transformation with respect to α and γ by maximizing the likelihood $P(h_\gamma(X^\alpha)|\alpha, \gamma, H, \lambda)$.

The procedure is described in Fig. **??**. For each warping index α and each string candidate H_n, we solve for the $\hat{\gamma}_{\alpha,H_n}$ which maximizes $P(h_\gamma(X^\alpha)|\alpha, \gamma, H_n, \lambda)$. Next, the warping index $\hat{\alpha}$ is selected so that $P(h_{\hat{\gamma}}(X^\alpha)|\alpha, \hat{\gamma}, H_n, \lambda)$ is maximized. Finally, the transformed observation vector $h_{\hat{\gamma}}(X^{\hat{\alpha}})$ is used in a second recognition pass to obtain the final recognized string.

B.2 Combining Frequency Warping, Spectral Shaping and Model Selection

We propose to extend the method outlined in Eq. (4) to include model selection. Transformation-based speaker normalization applied to both training and testing of acoustic models can only partially address the mismatch and variability issues. Condition-dependent training or maximum a posteriori adaptation can further help reduce variability and mismatch. For children speech, a family of acoustic models obtained from speakers of different age groups are used in parallel during decoding. If $\lambda^n,\ n = 1, .., N$ is a family of acoustic models the maximum likelihood criterion can be used to select the appropriate model in conjunction with optimize the parameters of the speaker normalization and model adaptation algorithms as follows

$$\{\hat{\alpha}, \hat{\gamma}, \hat{n}\} = \arg \max_{\{\alpha,\gamma,n\}} P(X^\alpha|\alpha, \gamma, \lambda^n_\gamma, H)\ . \tag{5}$$

By combining speaker normalization with parallel model selection further performance improvement can be obtained.

In the next section, we present ASR experiment results in the context of speaker normalization and adaptation parameters estimated from single utterances. In our case, $h_\gamma()$ is a simple linear bias applied to the means of the model distributions or the observation sequence [14], and $\lambda^n,\ n = 1, .., N$ is a family of age-group dependent acoustic models.

V. ASR Experiments

Speaker normalization experiments, where adaptation parameters for both frequency warping and model transformation were estimated from single utterances, included both connected digit and command and control tasks (databases in Table I). The specific form of the transformation used in the speaker normalization experiments was linear frequency warping followed by a single linear bias applied to the warped observation sequence

$$h_\gamma(X^\alpha(t)) = X^\alpha(t) - \gamma \tag{6}$$

where X^α(t) is the cepstrum observation vector at time t warped by $g_\alpha()$. To estimate the optimum γ it was assumed that only the highest scoring Gaussian in the mixture contributes to the likelihood computation thus simplifying the estimate

$$\hat{\gamma} = (\sum_t \frac{X^\alpha(t) - \mu_{j(t)}}{\sigma_{j(t)}})/(\sum_t \frac{1}{\sigma_{j(t)}}) \tag{7}$$

where $\mu_{j(t)}, \sigma_{j(t)}$ are the mean and variance of the most active Gaussian j in the mixture at time instant t.

[2]The inverse transformation has to be applied to the observations. For simplicity we use the same notation for transformations applied to either the observations or to the HMMs.

Model	Baseline %	Norm. %	Improv.%
Adult HMM	15.9	8.7	+45
Children HMM	6.7	4.9	+25
Adult+Child. HMM	7.6	5.6	+25

TABLE II

AVERAGE DIGIT ERROR RATE FOR CHILDREN SPEAKERS BEFORE AND AFTER SPEAKER NORMALIZATION.

The first experiment focussed on investigating age-dependent effects of the speaker normalization schemes. Various model conditions were considered – HMMs trained using utterances from adults and children with and without frequency warping and spectral shaping – during recognition. Results were obtained for the connected digits and command/control tasks and are shown in Fig. 9(a) and (b), respectively.

A. Connected Digit Recognition Task

Separate sets of acoustic models were trained using data from adults (DgtI corpus, labeled "Adult. HMM") and children (DgtII corpus, "Child. HMM"). A mixture of six Gaussians were used to model each of the three states of the context-dependent digit units. In Fig. 9(a), digit accuracy is shown for the test corpus DgtIII before and after combined frequency warping and model transformation for HMMs trained from both adult and children speech. The allowed range of frequency warping was from –20% to +12%; a total of 17 warping factors were examined during frequency warping. The relative error rate reduction due to speaker normalization was found to be up to 50%, and was greatest for young speakers under twelve years of age tested using models trained from adult speakers (dotted vs. dashed lines in Fig. 9(a)). After speaker normalization the recognition accuracy for children over nine years of age is comparable to that of adults.

A summary of the average results for all ages is given in Table II. It also includes the performance results for HMMs trained using roughly equal amounts data from the adult and children corpora DgtI and DgtII (labeled "Adult+Child. HMM"). Overall, digit error rate reduction is about 25-45% after speaker normalization procedures depending on the model condition. Note that on average only 3.5 digits were used to estimate the parameters of the frequency warping and the linear transformation.

Next, the effect of using age-dependent models was investigated. Due to data sparseness, age-dependent models were trained from corpus DgtII for only two speaker groups: ages 10-12 and 13-17 years. The maximum likelihood criterion (Eq. (5)) was used to select between the two models. After speaker normalization an additional 10% reduction in word error rate was achieved using age-dependent models. This indicates that further improvement in recognition performance might be possible by imposing such additional constraints on the ASR acoustic space.

B. Command Phrase Recognition Task

Again, two sets of context-independent phone models were trained using data obtained from the corpus SubwI ("Adult") and SubwII ("Child.") summarized in Table I. A mixture of 16 Gaussians were used to model each state of the the 40 context-independent (subword) English phone units. In addition, a garbage model (represented by 5 states, 16 Gaussian mixtures/state). and a silence model (single state, 32 Gaussians/state) were included. Fig. 9(b) shows word recognition accuracy as a function of age for for test data obtained from the CommI and CommII corpora. CommI

and CommII consist of 60 possible short phrases (85 words). Recognition was performed using a finite state grammar comprising the relevant phrases. The baseline recognition performance for the "Adult. HMM" (dotted line) decreases rapidly for speakers younger than twelve due to the increasing acoustic mismatch between the training and testing speaker populations. Similarly, recognition performance for the "Child. HMM" (dashed-dotted line) trails off for speaker ages 6-8 due to acoustic mismatch (no children younger than ten in training corpus SubwII) and increased acoustic variability for the 6-8 age group. The overall effects of speaker normalization are similar to those seen for the connected digit task.

C. Effects of Speaker Normalization on ASR Performance

Next the role of frequency warping and model transformation were investigated both when used independently and when combined. Since the results for connected digit and subword recognition tasks showed similar trends we focus on the former for the experiment considered here. Context-dependent ("head-body-tail") digit models were trained from equal amounts of data drawn from both adult and children speakers. A mixture of six Gaussians were used to model each state. The models were trained discriminatively using the generalized gradient descent algorithm (five iterations over the training data).

Fig. 10 shows the age-dependent effects of frequency warping and adaptation of model means. To observe overall trends due to warping and adaptation, results were also averaged across the age groups and summarized in Table III. The second row of the table, "Warp", refers to the warping algorithm of Sec. IV-B. The amount of linear frequency scaling ranged from -20% compression to +12% expansion with a total of 17 warping factors were allowed in this range. The third row of the table, "Bias", displays the recognition rate when a single linear bias is estimated for the whole utterance without the use of warping. The optimal bias vector $\hat{\gamma}$ maximizes $P(h_\gamma(X)|\gamma, \lambda, H)$, where H is the corresponding transcription obtained from a preliminary decoding pass. The fourth row of Table III, labeled "Warp+Bias", refers to combining frequency warping and bias estimation as in Eq. (4). Note that a separate bias vector $\hat{\gamma}_\alpha$ was computed and subtracted from each warped observation sequence X^α before the optimal warping index $\hat{\alpha}$ was selected.

Consider the age-dependent effects shown in Fig. 10. Overall, both warping and model adaptation contribute to improved performance across all age groups; their combined use provides the best performance. This is in agreement with our claim in Sec. IV-B that the combined optimization of both model transformation and frequency warping is important for obtaining a better match between the utterance and the model. Note that the reduction in error rate obtained by combining the warping and spectral shaping algorithms is approximately equal to the sum of the reduction in error rates when applying each of the adaptation procedures separately. Further, it is interesting to observe that the effect of frequency warping is more dominant for the younger age group (younger than 12 years) while model adaptation becomes dominant for the older group. Although the models were based on children speech, these results indicate the significance of variability within this group introduced by developmental changes (Sec. II).

On average, across age-groups, warping provides about 18.5% improvement and model adaptation, about 11.3% improvement (Table III). When combined, these methods provide about 28.7% improvement for this model condition. Frequency warping plays a greater role toward reducing mismatch, and hence error rates, for females than for males (7.4% vs. 22.5%). This is not surprising since the spectral parameter variability with age in female children was found to be larger compared to males (Sec. II). Model adaptation, on the other hand, appears to have a bigger effect on males (especially for younger children).

Algorithm	Male		Female		Both	
	Error%	Improvement%	Error%	Improvement%	Error%	Improvement%
Baseline	6.55	-	9.33	-	7.84	-
Bias	5.61	+14.3	8.77	+5.9	6.95	+11.3
Warp	6.07	+7.4	7.23	+22.5	6.39	+18.5
Warp+Bias	5.23	+21.7	6.69	+28.3	5.59	+28.7

TABLE III

SPEAKER NORMALIZATION EXPERIMENTS: INDIVIDUAL AND COMBINED EFFECTS OF LINEAR FREQUENCY WARPING ("WARP") AND MODEL ADAPTATION ("BIAS") ON DIGIT ERROR RATES. BASELINE HMMS WERE TRAINED FROM EQUAL AMOUNTS OF DATA FROM ADULT AND CHILDREN SPEAKERS.

Next let us consider the optimal warping factors estimated by the maximum likelihood approach. In Fig. 11, the average (across all digits strings) optimal warping factors $\hat{\alpha}$, as computed from the speaker normalization algorithm, are shown per speaker's age and gender for the adult HMMs. Note that $\alpha = 1$ corresponds to no warping (data matched with the adult model), while $\alpha = 0.8$ corresponds to 20% compression of the frequency scale. The elevated slope of the average warping factor curve for young male speakers corresponds to the rapid vocal tract growth during puberty. The plot can be compared with the warping factors computed via speech analysis in Fig. 7. Note, however, that the acoustic model used as a reference in Fig. 11 is trained from both adult male and female speech, while in Fig. 7(a) the warping factor is computed relative to adult male speech only (relative to adult female speech in Fig. 7(b)). Despite the difference in the reference model and in the algorithm used for computing the optimal warping factor, the results in Fig. 11 are consistent with the acoustic analysis results discussed in Sec. II and Sec. IV-A.

D. Bi-parametric Frequency Warping

Finally, we explore the usefulness of alternative frequency warping strategies compared to the simple linear frequency warping. As discussed in [8], the assumption that all formants scale linearly with the vocal tract length is correct only to the first order. For example, it was shown in Fig. II that different formant frequencies (F1, F2, F3) get scaled by different amounts, especially in the case of female speakers.

To account for different scaling factors for F1, F2 and F3 a simple bi-parametric frequency warping algorithm is proposed. The two warping function parameters are the low frequency α_L and high frequency α_H scaling factors. The warped frequency f_w is computed as

$$f_w = [(1 - f/f_{\max})\,\alpha_L + (f/f_{\max})\,\alpha_H]\,f \tag{8}$$

where $f_{\max}$ is the speech signal bandwidth.

A piece-wise linear warping function is used, where different amount of warping is applied to each of two frequency bands. The values of α_L and α_H are determined by exhaustively searching over a grid of possible scale factor values so that the likelihood (see Eq. (4)) is maximized. A single utterance was used to estimate the scaling factors. The speaker normalization function was tested on the DgtIII corpus (Fig. 9(a)) using a set of 40 possible (α_L, α_H) combinations, ranging from –20% to +12% under the beam constraint $|\alpha_L - \alpha_H| \leq 0.06$. An *additional* 3-5% reduction in error rate (mostly for female speakers) was achieved when using the bi-parametric vs. linear frequency warping function. The average low and high frequency scaling factors computed from the speaker

normalization algorithm display similar trends to the formant scaling factors for F1 and F2, F3 computed in [8]. On average, it was found that $|\alpha_L - 1| > |\alpha_H - 1|$, i.e., the low frequency band (corresponding roughly to F1) gets expanded or compressed more that the high frequency band (F2, F3), especially for female speakers.

VI. Summary

Children's speech is different from adult's speech in terms of both magnitude and variability of acoustic and linguistic correlates. As a result, the acoustic space of children's speech is much larger and is characterized by highly overlapping phonemic classes. Developmental changes, especially vocal tract growth, contribute to variability in spectral and temporal parameters of the speech signal of children. These factors pose challenges to automatic speech recognition. Simple speaker normalization procedures were shown to reduce the acoustic variability and mismatch, both within and between the children and the adult acoustic spaces.

With the combined use of linear frequency warping, model adaptation, and age-dependent acoustic modeling, it was shown that recognition performance could be improved by up to 55% for children speakers, using just a single utterance for estimating normalization parameters. The usefulness of bi-parametric warping function was investigated as an alternative to linear frequency warping and was shown to provide some additional improvements in ASR performance. Furthermore, the use of phoneme-dependent scaling factors was shown to further reduce spectral mismatch between the original and target group of speakers over using a global scaling factor. In practice, however, these gains can be rather limited because of the great number of scaling factors that have to be estimated. A promising alternative is to compute class-dependent scaling factors using three or four broad phonemic classes.

The frequency warping algorithm described in this paper, although simple, still requires two recognition passes. There is a need for developing rapid normalization and adaptation procedures to enable widespread adoption of these techniques in live voice-enabled systems for children. Further work is also needed to investigate non-linear warping and phonemic-class dependent frequency warping algorithms. The results of this paper nevertheless demonstrate that speech recognition for children speakers is viable and good ASR performance can be obtained even under mismatched conditions.

VII. Acknowledgments

The authors are grateful to Dr. Richard Rose at AT&T Labs-Research and Dr. Sungbok Lee at Bell Labs, Lucent Technologies for discussions and help related to this work. Most of this work was done when the authors were with AT&T Labs-Research.

References

[1] D. C. Burnett and M. Fanty, "Rapid unsupervised adaptation to children's speech on a connected-digit task," in *Proc. ICSLP*, Oct. 1996.

[2] S. Das, D. Nix and M. Picheny, "Improvements in children's speech recognition performance", in *Proc. ICASSP*, pp. 433–436, 1998.

[3] S. Eguchi and I. J. Hirsh, "Development of speech sounds in children," in *Acta. Otolaryng.*, Suppl. vol. 257, 1969.

[4] U. G. Goldstein, "An articulatory model for the vocal tracts of growing children," *Ph.D. Thesis*, MIT, 1980.

[5] J. Hillenbrand, L. Getty, M. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. 97, pp. 3099–3111, 1995.

[6] R. D. Kent, "Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies," *JSHR*, vol. 19, pp. 421–447, 1976.
[7] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. ICASSP*, pp. 353–356, May 1996.
[8] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *J. Acoust. Soc. Am.*, vol. 105, pp. 1455–1468, Mar. 1999.
[9] J. D. Miller, S. Lee, R. M. Uchanski, A. F. Heidbreder, B. B. Richman and J. Tadlock, "Creation of two children's speech databases," in *Proc. ICASSP*, pp. 849–852, 1996.
[10] J. Mostow, A. G. Hauptmann, and S. F. Roth, "Demonstration of a reading coach that listens," *Proceedings of the ACM Symposium on User Interface Software and Technology*, pp. 77–78, 1995.
[11] S. Narayanan and A. Potamianos, "Creating conversational interfaces for children," in *IEEE Trans. Speech and Audio Proc.*, To appear, 2002.
[12] S. Palethorpe, R. Wales, J. Clark and T. Senserrick, "Vowel Classification in Children," *J. Acoust. Soc. Am.*, vol. 100, pp. 3843–3851, Dec. 1996.
[13] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," J. Acoust. Soc. Am. 24., pp. 175–184, 1952.
[14] A. Potamianos and R. C. Rose, "On combining frequency warping and spectral shaping in HMM-based speech recognition," in *Proc. ICASSP*, Apr. 1997.
[15] A. Potamianos, S. Narayanan, and S. Lee, "Automatic speech recognition for children," in *Proc. EuroSpeech*, vol. 5, (Rhodes, Greece), pp. 2371–2374, Sept. 1997.
[16] M. Russell, B. Brown, A. Skilling, R. Series, J. Wallace, B. Bonham, and P. Barker, "Applications of automatic speech recognition to speech and language development in young children," in *Proc. ICSLP*, Philadelphia, PA, Oct. 1996.
[17] B. L. Smith, "Relationships between duration and temporal variability in children's speech," J. Acoust. Soc. Am., vol. 91, pp. 2165–2174, 1992.
[18] E. F. Strommen and F. S. Frome, "Talking back to big bird: Preschool users and a simple speech recognition system," *Educational Technology Research and Development*, vol. 41, pp. 5–16, 1993.
[19] J. G. Wilpon and C. N. Jacobsen, "A study of automatic speech recognition for children and the elderly," in *Proc. ICASSP*, pp. 349–352, May 1996.
[20] V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T. Hazen, and L. Hetherington, "Jupiter: A telephone-based conversational interface for weather information," *IEEE Trans. Speech Audio Processing*, vol. 8, Jan. 2000, pp. 85-96.

List of Tables

List of Figures

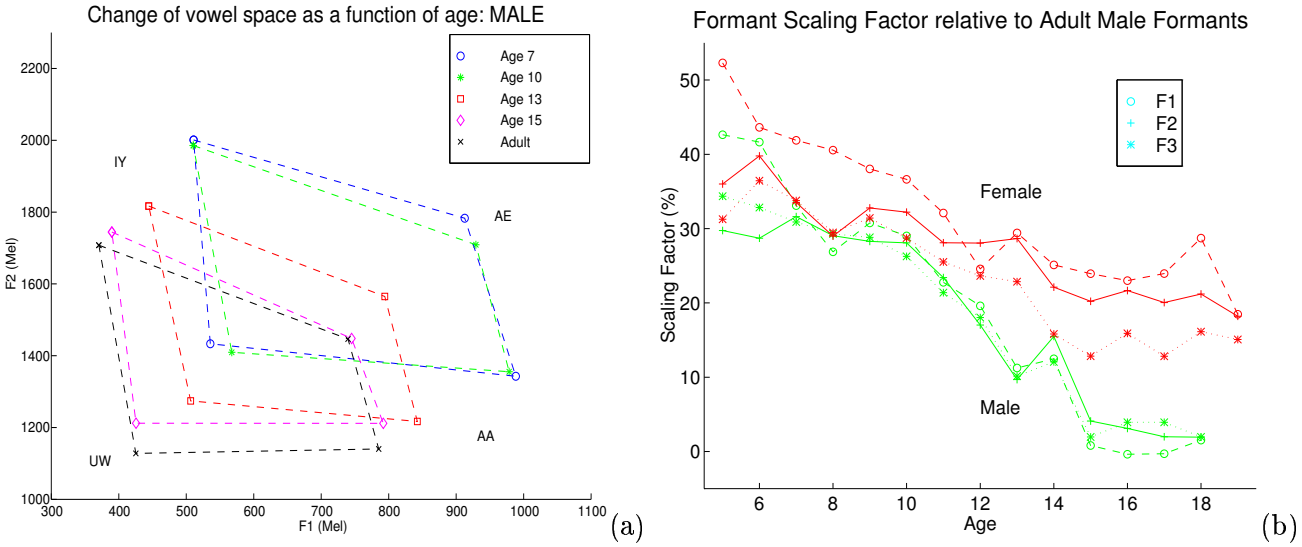

Fig. 1. (a) Changes in F1-F2 vowel space as a function of age. The vowel space boundaries are marked by average formant frequency values for the four point vowels /AA, IY, UW, AE/ for the age groups: 7, 10, 13, 15 and adults. (b) Scaling factor variation in first three formant frequencies with respect to age for vowels of male and female children. Scaling was with respect to average values for adult males.

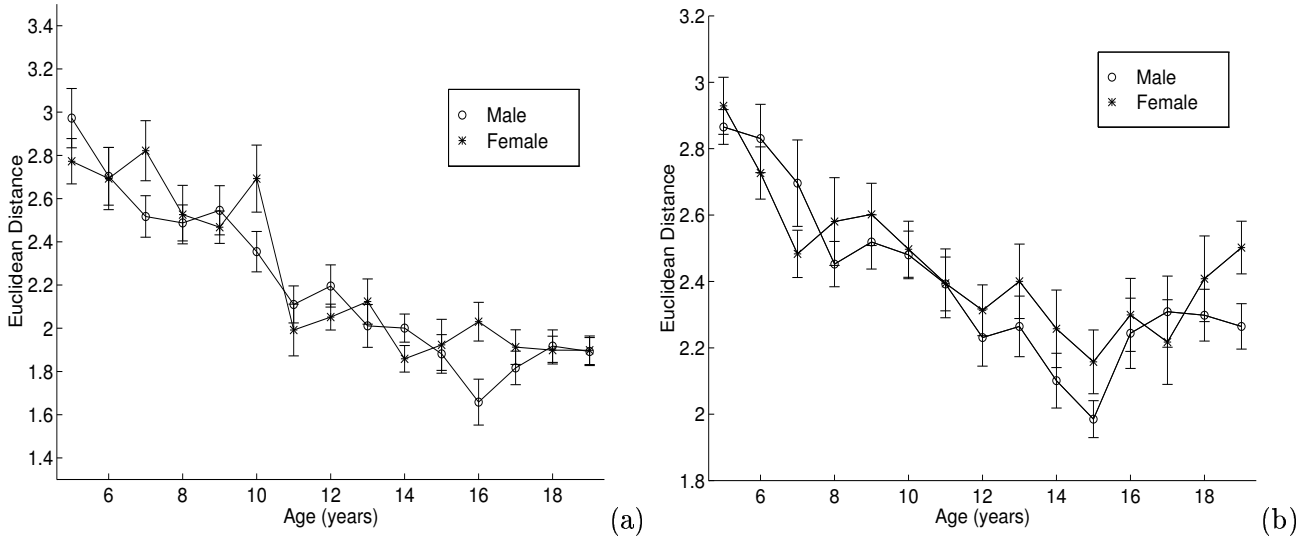

Fig. 2. Intra-speaker variability as a function of age: (a) Mean cepstral distance between the two repetitions of the same vowels and (b) Mean cepstral distance between the first– and second–half segments within the same vowel realization.

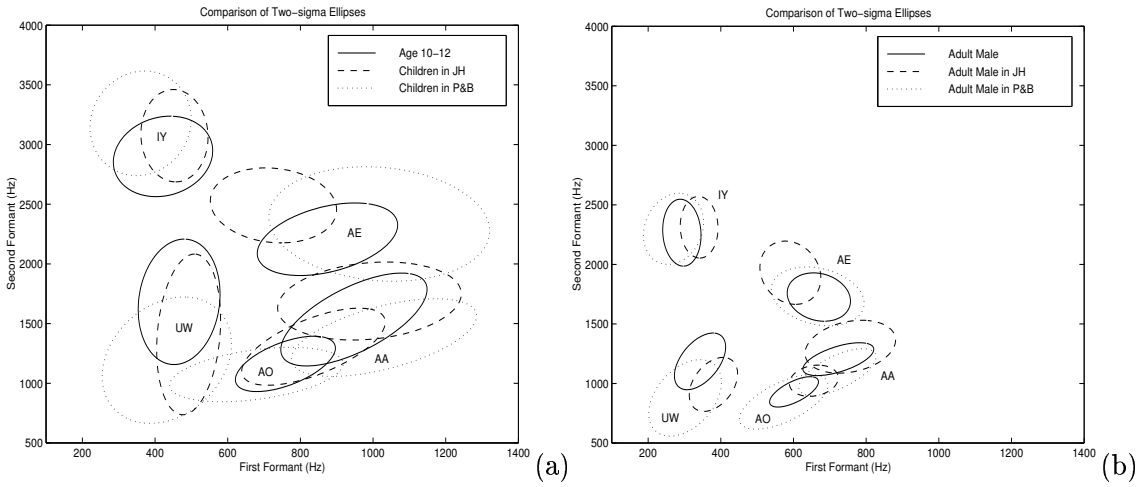

Fig. 3. (a),(b) Comparison of formant data of children (ages 10 through 12) and adults : data from Lee et al [8], Hillenbrand et al [5], Peterson and Barney [12].

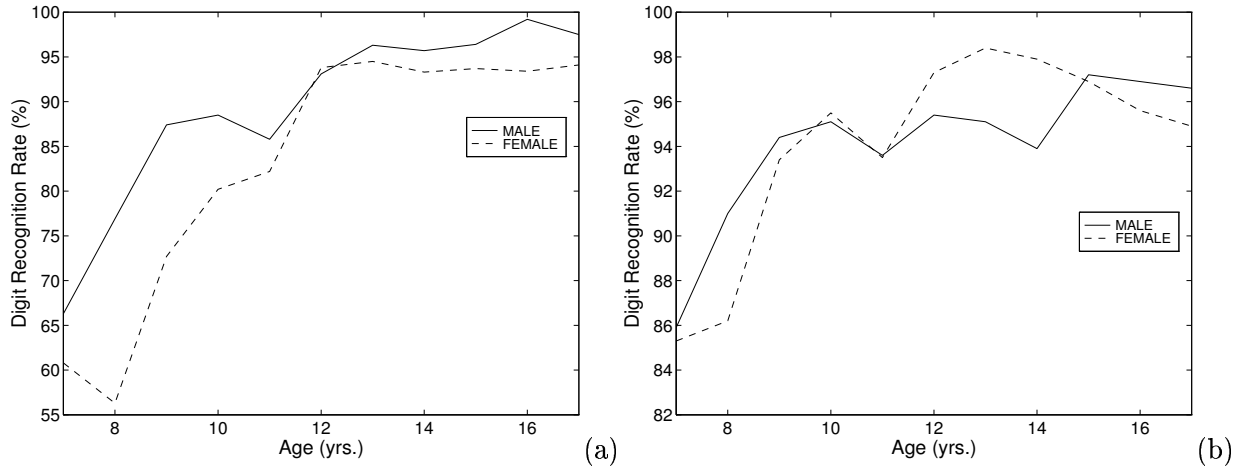

Fig. 4. Word recognition rate as a function of age and gender using acoustic models trained from (a) adult and (b) children speakers.

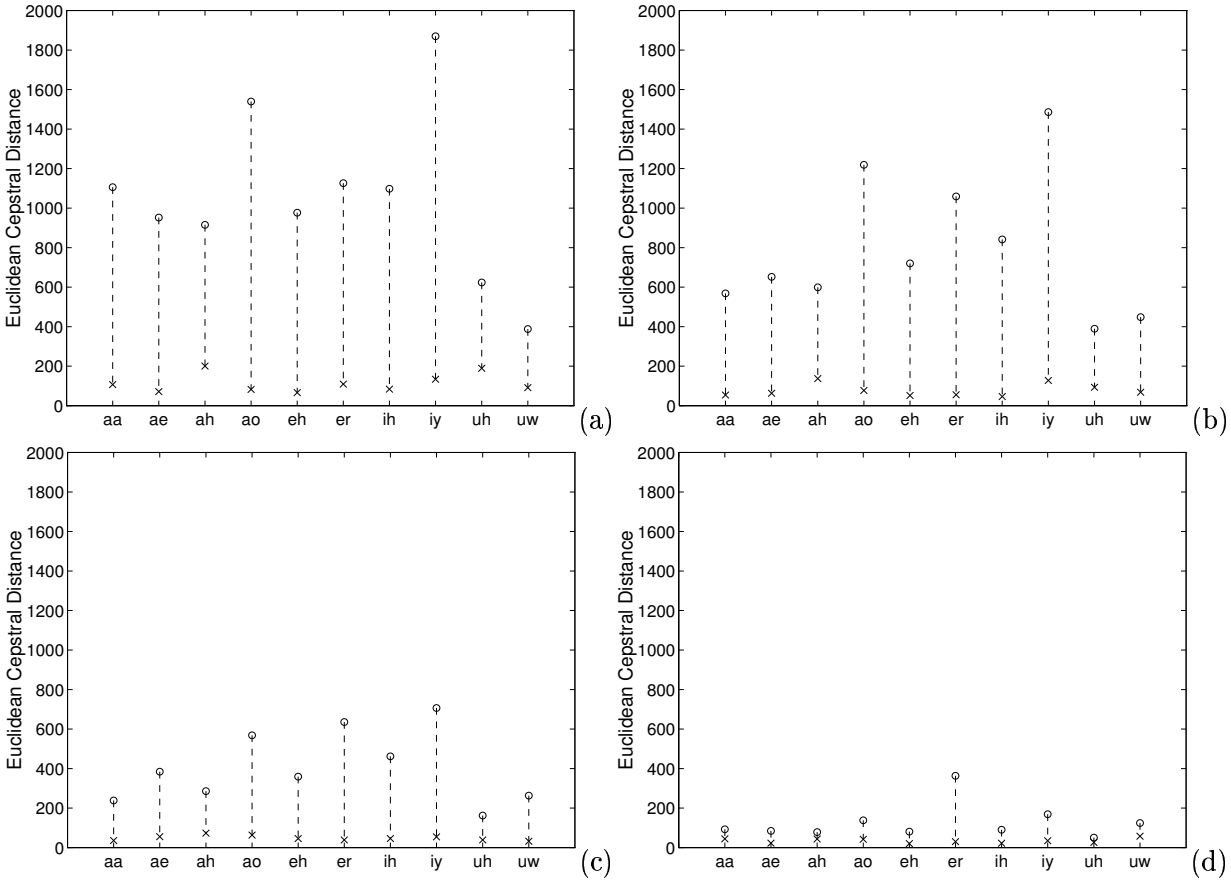

Fig. 5. (a)-(d) Average Euclidean cepstrum distance between male children speakers and male adult speakers before (o) and after (x) frequency warping for all ten monophthongal vowels. The optimal scaling factors were selected for each phoneme, speaker age, and gender. Averages for age groups 5-7 (a), 8-10 (b), 11-13 (c) and 14-16 (d) years are shown.

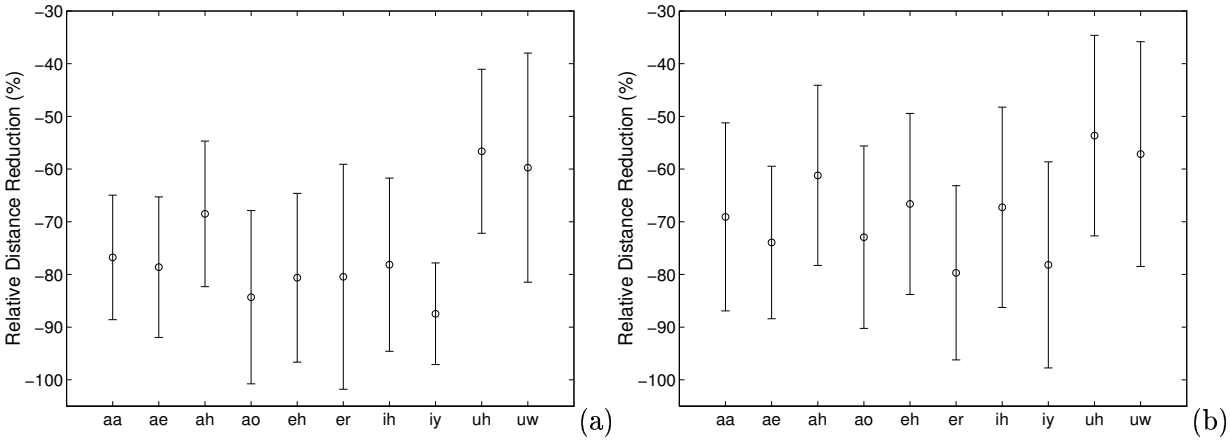

Fig. 6. (a),(b) Percent distance reduction due to frequency warping when *scaling factors and distance reduction are computed on an per utterance basis.* Mean and standard deviation of distance reduction (error bars) is displayed for age groups 5-7 (a) and 11-13 years (b).

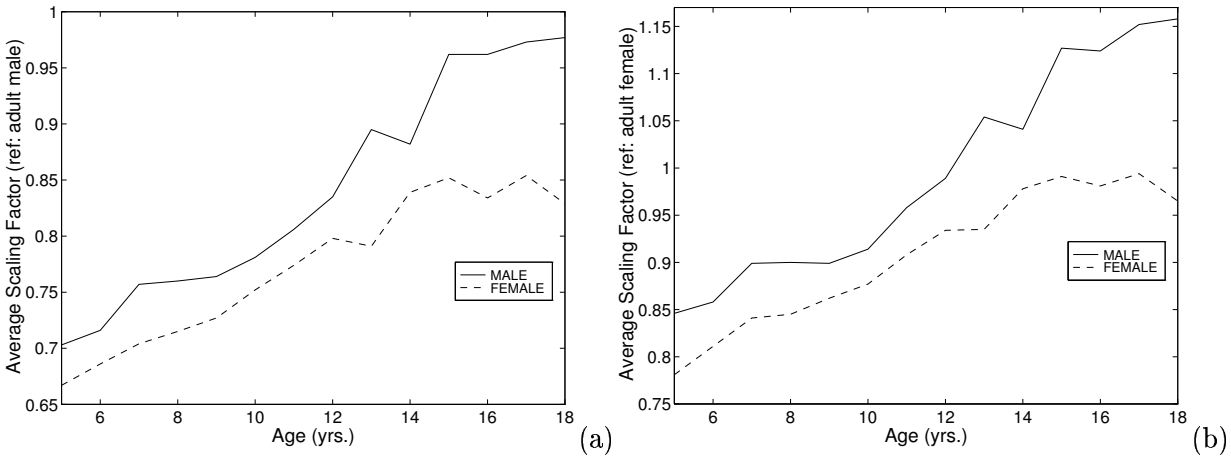

Fig. 7. Optimal warping factors averaged over 10 monophtongal vowels as a function of age. Warping factors are computed by minimizing the Euclidean distance between the reference adult spectral envelope (corresponding to scale factor 1.00) and the warped average spectral envelope for each age and each vowel. Reference is adult male in (a) and adult female in (b). Note that the y-axis scales in (a) and (b) are different.

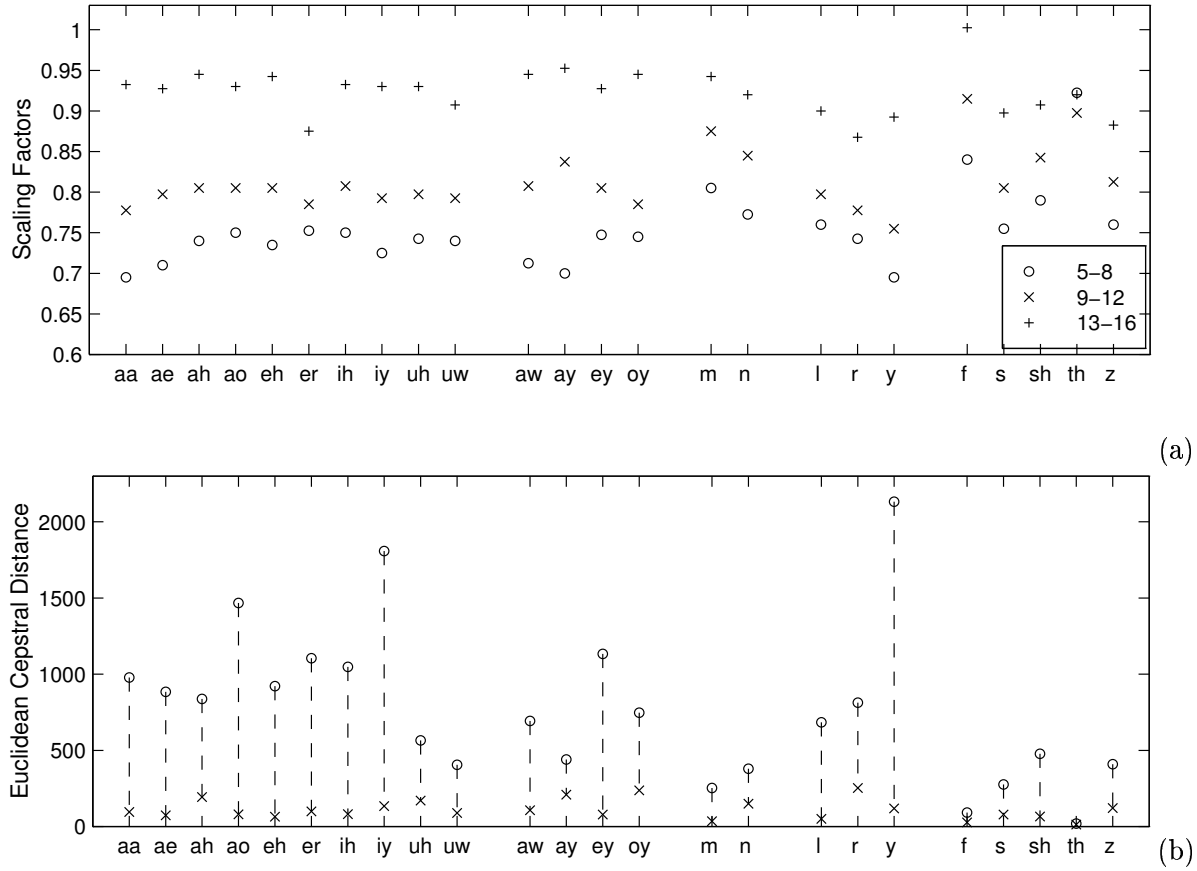

Fig. 8. (a) Optimal scaling factors for vowels, nasals, glides and fricatives for male children ages 5-8 (o), 9-12 (x), 13-16(+) (reference male adult speakers). (b) Average Euclidean cepstrum distance between children male speakers ages 5-8 and adult male speakers before (o) and after (x) frequency warping.

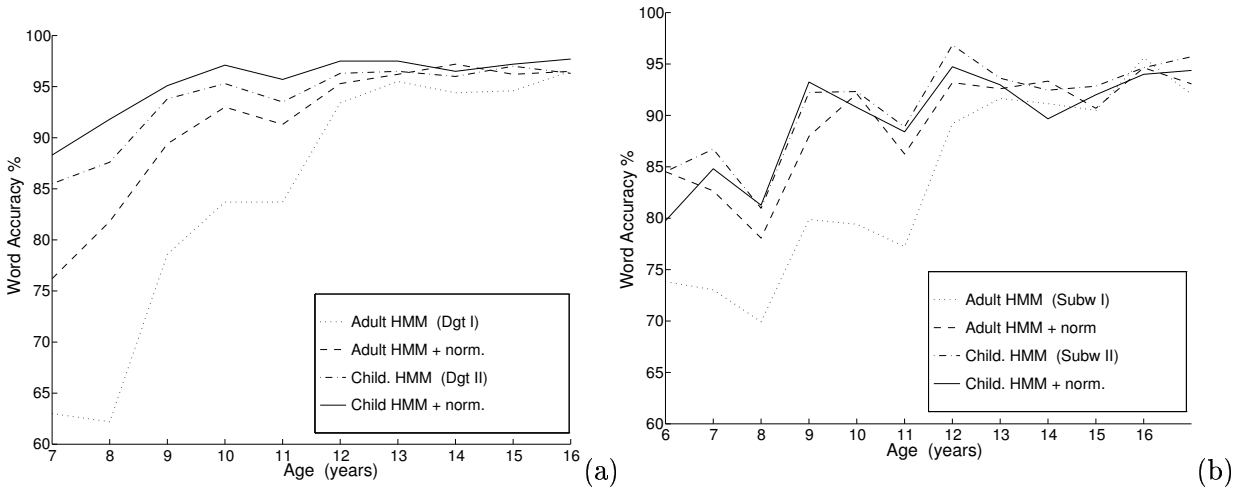

Fig. 9. Word accuracy (%) vs. speaker's age using HMMs trained from children or adult speakers before and after speaker normalization algorithms were applied. Test databases: (a) Connected digits (DgtIII), (b) Command and control phrases (Comm). Results with adult HMMs without normalization (dotted), with normalization (dashed), child HMMs without normalization (dot-dashed) and with normalization(solid).

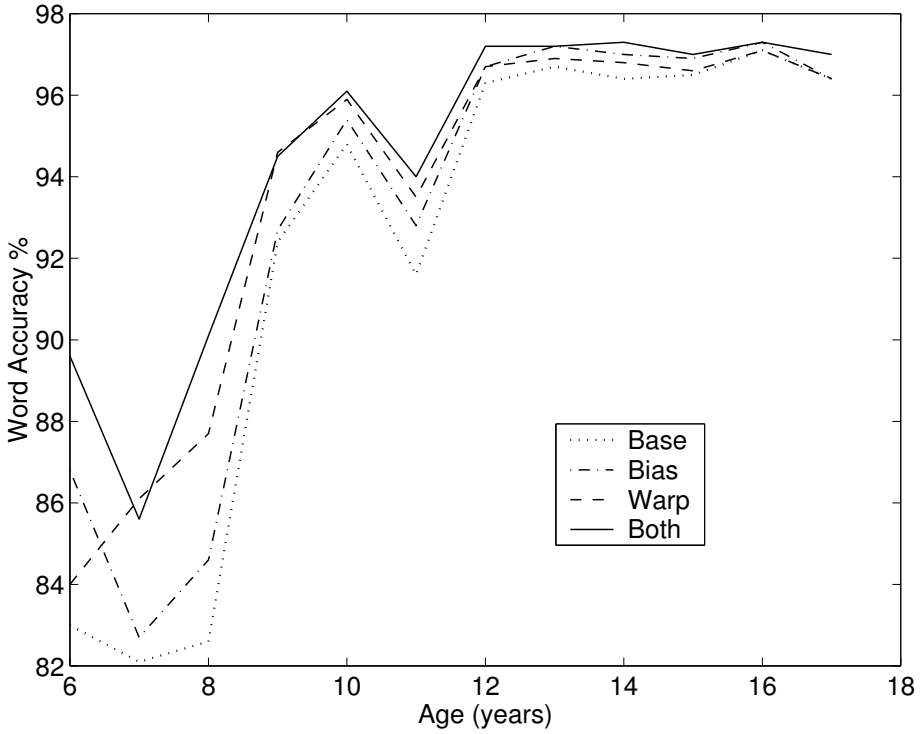

Fig. 10. Effects of speaker normalization on word accuracy as a function of age for digit recognition using using acoustic models trained from adults and children: baseline (dotted), adaptation with transformation of model means (dot-dashed), linear frequency warping (dashed), combined frequency warping and model adaptation (solid).

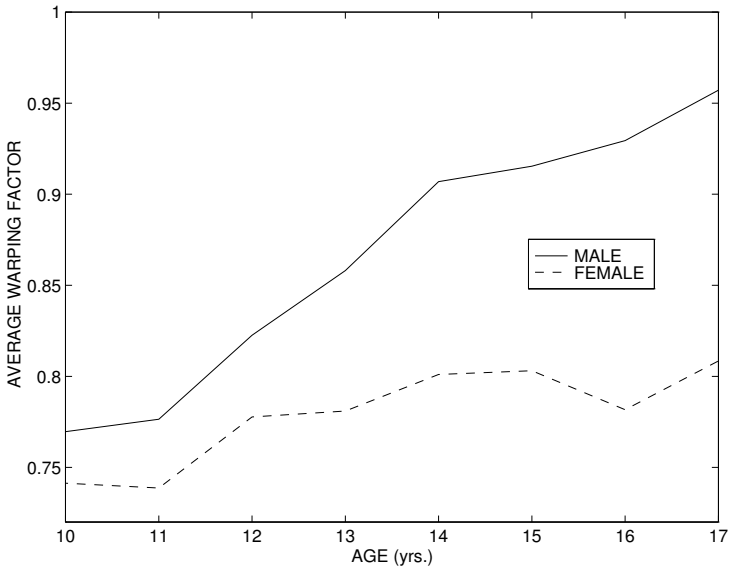

Fig. 11. Average warping factors per age and gender for the connected digit recognition task computed via maximum likelihood frequency warping using an adult speech HMM.