AUTOMATIC SPEECH RECOGNITION FOR CHILDREN

*Alexandros Potamianos, Shrikanth Narayanan and Sungbok Lee**

AT&T Labs–Research, 180 Park Ave, P.O. Box 971, Florham Park, NJ 07932-0971, U.S.A.
email: {potam,shri,sungbok}@research.att.com

ABSTRACT

In this paper, the acoustic and linguistic characteristics of children speech are investigated in the context of automatic speech recognition. Acoustic variability is identified as a major hurdle in building high performance ASR applications for children. A simple speaker normalization algorithm combining frequency warping and spectral shaping introduced in [5] is shown to reduce acoustic variability and significantly improve recognition performance for children speakers (by 25–45%). Age-dependent acoustic modeling further reduces word error rate by 10%. Piece-wise linear and phoneme-dependent frequency warping algorithms are proposed for reducing acoustic mismatch between the children and adult acoustic spaces.

1. INTRODUCTION

Automatic speech recognition (ASR) for children speakers is a challenging problem with many potential applications. Although a significant amount of literature exists on comparative analysis of the acoustic and linguistic characteristics of children with those of adults [2, 4], our understanding of how such differences affect speech recognition performance is limited. In this paper, we present new results of ongoing efforts to improve the performance of speech recognition for children speakers [6].

The acoustic and linguistic characteristics of children's speech change rapidly as a function of age and are widely different from those of adults. 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 features such as prosody. 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*. [2, 4]. 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. A detailed report of the temporal and spectral characteristics of children's speech as a function of age can be found in [4].

There are several implications that the acoustic differences mentioned above have on speech recognition for children. The main goal of the ASR feature extraction stage is to decompose the speaker-dependent information (e.g., pitch) from the phoneme-dependent information (e.g., formants) and retain the latter. This task is more difficult for children voices because the fundamental frequency and the formant bandwidths are of comparable magnitude. Moreover, for telephone speech, a large spectral slice containing the high-frequency formants is lost due to band-limiting. 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.

A major hurdle in acoustic modeling of children speakers is spectral and temporal variability in children 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 [4]. The combination of a large acoustic parameter range and increased acoustic variability can seriously degrade ASR performance. In Section 2, speaker normalization procedures and age-dependent acoustic modeling are used to reduce variability and increase resolution between classes.

Other important issues in ASR for children, not tackled in this paper, is the spontaneity and greater linguistic variability of children's speech (that creates large amounts of extraneous speech) and associated ASR interface issues. In general, the modality of child-machine interaction using spontaneous speech is surprisingly different than that of adults and needs to be studied in detail.

As discussed above the large range of values and the increased variability of acoustic and temporal features in children speakers pose a challenging problem for ASR. There are certain other characteristics of children speech, however, that could be potentially advantageous for ASR. In [4], young children are shown to be less skilled in coarticulation, and display longer durations especially across word boundaries. This signifies that simpler (context-independent) acoustic models can be used for certain ASR tasks. Further, children tend to exaggerate newly acquired or recently mastered skills. This "overshooting" tendency is clear in the measurements of spectral and temporal parameters in [4]. For example, five-year old children show a tendency to over-elongate certain vowels (/iy/, /aa/, /ae/, /uw/) to differentiate them from their confusable "short" counterparts (/ih/, /ah/, /eh/, /uh/). Such spectral or temporal patterns could potentially improve ASR performance provided that the classifier and feature extraction system are able to efficiently incorporate such cues.

Overall, current ASR systems are unable to cope with the increased degree of variability and spontaneity in children speech, due to lack of general normalization algorithms. In Section 2, we investigate how these sources of acoustic mismatch and variability in children speech affect speech recognition performance. A speaker normalization procedure that combines spectral shaping and frequency warping is implemented that reduces recognition error rate up to 45%. In Section 3, bi-parametric

*Sungbok Lee is currently with the Central Institute for the Deaf, St. Louis, MO.

and phoneme-dependent mappings from the children to the adult ASR acoustic feature space are investigated.

2. ACOUSTIC MODELING

In this section, we evaluate speech recognition performance as a function of speaker's age. Acoustic mismatch and variability are identified as the major contributors to performance degradation. Speaker normalization, model adaptation and age-dependent models are used to reduce variability and constrain the acoustic space of children speakers. Substantial performance gains are achieved.

Acoustic models were trained from utterances collected over the public switched telephone network from either adult or children speakers. A summary of the training and testing databases is provided in Table 1. In Fig. 1(a), word accuracy as a function of age is shown for models trained from an adult speaker population (DgtI) labeled "ADLT HMM" and from a children speaker population (DgtII) labeled "CHLD HMM" for a connected digit recognition task on corpus DgtIII. For both matched and (especially for) mismatched training and testing conditions recognition performance decreases substantially for young children. Performance reaches adult levels approximately around thirteen or fourteen years of age, which agrees with the observation in [4] that by the age of fourteen *both* the mean and standard deviation of most acoustic characteristics have reached adult levels. Overall, recognition performance for children speakers *up to four times worse* than for adults depending on the speaker's age. For mismatched training and testing conditions ("ADLT HMM") word error rate is approximately two to three times higher than for matched conditions ("CHLD HMM"). We have also observed (results not shown here) that a relatively small improvement is achieved by using context-dependent (vs. context-independent) model units, which agrees with the observation in [4] that young children (ages 5-12) have not fully developed their co-articulation skills.

The major reason for performance degradation for younger speakers is acoustic mismatch between the training and testing data, increased acoustic variability and the large range of acoustic parameters (as discussed in Section 1). Next, we attempt to improve recognition performance by attacking each one of these problems. Speaker normalization and model adaptation is used to reduce the mismatch and variability, and age-dependent models are used to constrain the acoustic space under consideration.

2.1. Speaker Normalization and Adaptation

In [5], a parametric linear transformation of the HMM models and a parametric frequency warping of the input utterance were combined under a single statistical framework. Next, we outline the joint normalization and adaptation procedure, and propose an extension for a family of HMMs. The goal is to improve recognition performance by improving the match between HMMs and test utterances, and by reducing acoustic variability of the HMMs.

Name	Speaker Population	Content	No. of speakers	No. of strings
DgtI	Adults	digits	3026	4781
DgtII	10-17 yrs.	digits	1234	5767
SubwI	Adults	phrases	242	12144
SubwII	10-17 yrs.	phrases	1234	14267
DgtIII	6-17 yrs.	digits	501	2656
CommI	6-17 yrs.	commands	501	3554
CommII	10-17 yrs.	commands	1234	7436

Table 1: Training and testing databases.

The frequency warping approach to speaker normalization compensates mostly for inter-speaker vocal tract length variability by linear warping of the frequency axis by a factor α [3]. Frequency warping is implemented in the mel-frequency filterbank front-end by linear scaling of the spacing and bandwidth of the filters. 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. Let X^{α} denote the sequence of cepstrum observation vectors warped by a linear frequency warping function. If λ denotes the parameters of the HMM model, then the optimal warping factor is defined as

$$\hat{\alpha} = \arg\max_{\alpha} P(X^{\alpha}|\alpha, \lambda, H) \qquad (1)$$

where H is a decoded string obtained from an initial recognition pass. The selected observation vector sequence $X^{\hat{\alpha}}$ is decoded in a second recognition pass to obtain the recognized string.

There is a large class of maximum likelihood based model adaptation procedures that can be described as parametric transformations of the HMM model or the observation sequence. For these procedures, we let $\lambda_{\gamma} = h_{\gamma}(\lambda)$ denote the model obtained by a parametric linear transformation $h_{\gamma}()$. The optimal parameters of the linear transformation $\hat{\gamma}$ and the frequency warping $\hat{\alpha}$ can be simultaneously estimated. Further, if $\lambda^n,\ n = 1, .., N$ is a family of acoustic models the maximum likelihood criterion can be used to select the appropriate model and also optimize the parameters of the speaker normalization and model adaptation algorithms as follows

$$\{\hat{\alpha}, \hat{\gamma}, \hat{n}\} = \arg\max_{\{\alpha,\gamma,n\}} P(X^{\alpha}|\alpha, \gamma, \lambda^n_{\gamma}, H)\ . \qquad (2)$$

The potential of this class of procedures was investigated in the context of speaker adaptation 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 [5], and $\lambda^n,\ n = 1, .., N$ is a family of age-group dependent acoustic models.

2.2. Experimental Results

Next, speaker normalization and age-dependent acoustic modeling techniques are applied to the connected digits, and command and control recognition tasks.

Digit Recognition Task:

Acoustic models were trained from corpus DgtI (labeled "ADLT HMM") and DgtII ("CHLD HMM") in Table 1. A mixture of 6 Gaussians were used to model each state of the context-dependent digit units. In Fig. 1(a), the digit accuracy is shown for the test corpus DgtIII before and after speaker normalization. Results for the HMMs trained from adult and children speaker populations are shown. The allowed range of formant frequency scaling was from –20% to +12% and a total of 17 warping factors were examined during frequency warping. The error rate reduction due to speaker normalization is shown to be up to 50%, and is greater for young speakers under twelve years of age and for the mismatched "ADLT HMM" trained from adult speakers (dotted vs. dashed line). After speaker normalization the recognition accuracy for children speakers over 9 years of age is comparable to that of adults. The summary of the cumulative results

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

Table 2: Digit error rate for children speakers before and after speaker normalization.

for all ages is given in Table 2. In addition, the performance of an HMM trained from data (equally) mixed from the adult and children corpora DgtI and DgtII is shown (labeled "Cld+Adlt HMM"). Overall, digit error rate reduction by 25-45% was achieved when using the speaker normalization procedure despite the fact that on the average only 3.5 digits were used to estimate the parameters of the frequency warping and the linear transformation.

Age-dependent models were trained from corpus DgtII from two speaker groups: ages 10-12 and 13-17. The maximum likelihood criterion (Eq. (2)) 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. Further improvement in recognition performance might be possible by imposing additional constraints on the ASR acoustic space.

Command Phrase Recognition Task:

HMMs were trained from the corpus SubwI ("ADLT") and SubwII ("CHLD") in Table 1. A mixture of 16 Gaussians were used to model each state of the the 40 context-independent (subword) English phone units. In Fig. 1(b), (c), word recognition accuracy is shown as a function of age for test data from the CommI and CommII corpora. CommI and CommII consist of 10 possible phrases (16 words) and 50 phrases (68 words), respectively. Recognition was performed using a finite state grammar comprising the relevant phrases for each database. The baseline recognition performance for the "ADLT 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 "CHLD 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. Similarly to the digit recognition task, speaker normalization helps significantly to bridge the gap in performance between the models trained from adult and from children speaker populations. However, adult recognition accuracy levels are still not reached for the younger age group (6-9 years), suggesting that normalization strategies more sophisticated than simple linear frequency warping may be needed.

3. ISSUES IN SPEAKER NORMALIZATION

In the previous section, a simple linear frequency warping function was used to map from the children into the adult acoustic space. As discussed in [4], the assumption that all formants scale linearly with the vocal tract length is correct only to the first order. In reality, the average formant frequencies (F1, F2, F3) get scaled by different amounts, especially for female speakers. Further, the amount of formant scaling between children and adult speakers is phoneme-dependent. Next, bi-parametric and phoneme-dependent frequency warping functions are investigated for speaker normalization.

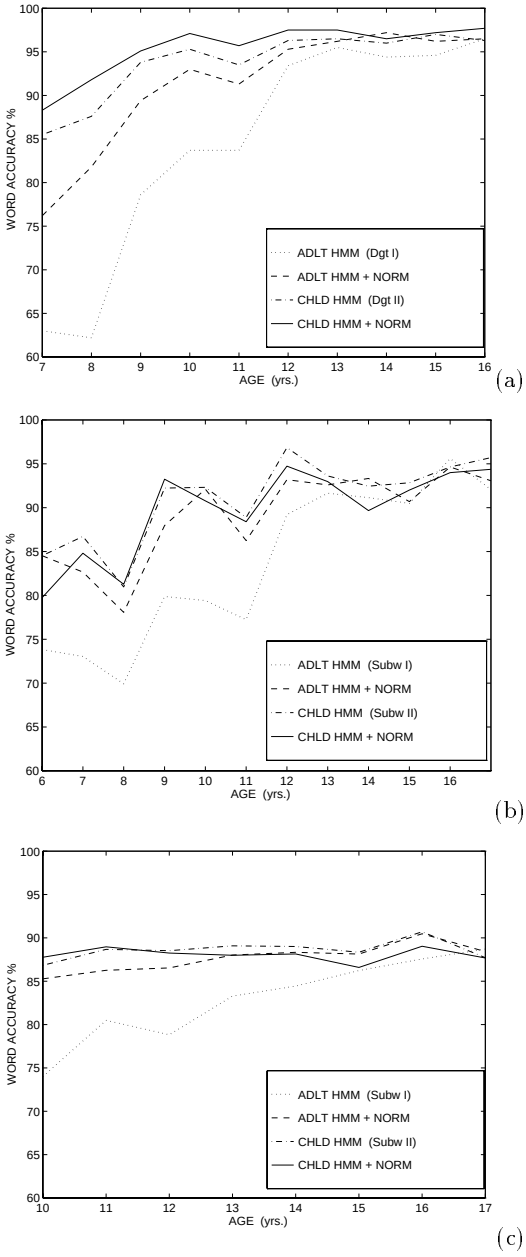

Figure 1: Word accuracy (%) vs. speaker's age using HMMs trained from children or adult speakers before and after the speaker normalization algorithm is applied. Test databases: (a) DgtIII, (b) CommI, and (c) CommII.

3.1. Bi-parametric Frequency Warping

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

$$f_w = [(1 - f/f_{\max})\ \alpha_L + (f/f_{\max})\ \alpha_H]\ f \qquad (3)$$

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

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. (2)) is maximized. A *single utterance* is used to estimate the scaling factors. The speaker normalization function was tested on the DgtIII corpus (Fig. 1(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 [4]. On the average $|\alpha_L - 1| > |\alpha_H - 1|$, 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.

3.2. Phoneme-dependent Frequency Warping

In this section, we investigate the need for a speaker normalization procedure that uses phoneme-dependent scaling factors. Further, the effectiveness of the speaker normalization algorithm is evaluated *for each phoneme* by comparing the spectral distances between the children utterances and the adult target ones before and after frequency warping is performed.

In Fig. 2(a), the optimal scaling factors (corresponding to the minimum Euclidean cepstrum distance between the normalized children utterances and the adult target ones) for male children of various age groups relative to adult male speakers are shown for monophthongal vowels, diphthongs, nasals, glides and fricatives. The scaling factors are computed for the average spectral envelope of all instances of the specified phoneme between speakers in the age groups 5-8, 9-12 and 13-16 years and adult male speakers. The C.I.D. high-quality microphone children database was used for that purpose (for details see [4]). The inter-phonemic scale factor variability for each of the age groups is relatively small and is greatest for the 5-8 age group. Scaling factors typically are more phoneme-dependent for diphthongs, glides and nasals than for vowels. For fricatives and nasals there is less age-dependent spectral change than for the rest of the phonemic classes. Note, that for all phonemes the scaling factors show similar trends as a function of age and thus using a phoneme-independent scaling factor is a valid approximation.

In Fig. 2(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. 2(a)) is shown to be very efficient in reducing acoustic mismatch between the young children and adult speakers for most phonemic classes.

4. CONCLUSIONS

Children speech is quite different from adult speech both in terms of absolute values and variability of acoustic and linguistic correlates. As a result the children acoustic space is large and with highly overlapping phonemic classes. Simple speaker normalization procedures were investigated that reduce acoustic variability and mismatch between the children and the adult acoustic spaces. The proposed linear warping speaker normalization, spectral shaping adaptation, and age-dependent acoustic modeling improved recognition performance up to 55% for children speakers, using a single utterance for adaptation. Finally, bi-parametric and phoneme-dependent warping functions were investigated as alternatives to linear frequency warping.

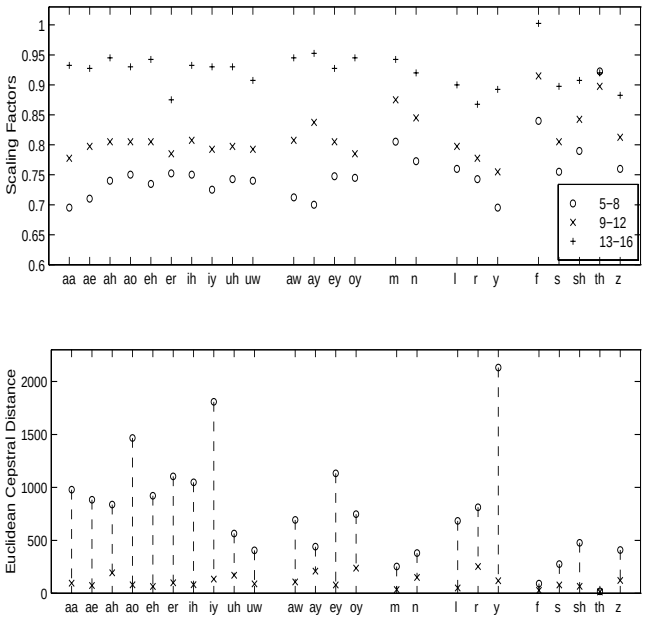

Figure 2: (a) Optimal scaling factors for vowels, nasals, glides and fricatives for male children ages 5-8 (o), 9-12 (x), 13-16(+) (reference male adult speakers). (b) Average Euclidean cepstrum distance between children male speakers ages 5-8 and adult male speakers before (o) and after (x) frequency warping.

5. ACKNOWLEDGMENTS

The authors would like to thank Rick Rose and Jay Wilpon for useful discussions and their help relating to this work.

6. 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] R. D. Kent, "Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies," *JSHR*, vol. 19, pp. 421–447, 1976.

[3] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. ICASSP*, pp. 353–356, May 1996.

[4] S. Lee, A. Potamianos, and S. Narayanan, "Analysis of children's speech: Duration, Pitch and Formant", in *Proc. EUROSPEECH 97*.

[5] A. Potamianos and R. C. Rose, "On combining frequency warping and spectral shaping in HMM-based speech recognition," in *Proc. ICASSP*, Apr. 1997.

[6] 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.