Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2010, Article ID 318785, 15 pages
doi:10.1155/2010/318785

Research Article

Exploring the Effect of Differences in the Acoustic Correlates of Adults' and Children's Speech in the Context of Automatic Speech Recognition

Shweta Ghai and Rohit Sinha

Department of Electronics and Communication Engineering, Indian Institute of Technology Guwahati, Guwahati 781039, India

Correspondence should be addressed to Shweta Ghai, shweta@iitg.ernet.in

Received 1 June 2009; Revised 30 October 2009; Accepted 25 January 2010

Academic Editor: Elmar Nöth

This work explores the effect of mismatches between adults' and children's speech due to differences in various acoustic correlates on the automatic speech recognition performance under mismatched conditions. The different correlates studied in this work include the pitch, the speaking rate, the glottal parameters (open quotient, return quotient, and speech quotient), and the formant frequencies. An effort is made to quantify the effect of these correlates by explicitly normalizing each of them using the already existing techniques available in literature. Our initial study done on a connected digit recognition task shows that among these parameters only the formant frequencies, the pitch, and the speaking rate affect the automatic speech recognition performance. Significant improvements are obtained in the performance with normalization of these three parameters. With combined normalization of the pitch, the speaking rate, and the formant frequencies, 80% and 70% relative improvements are obtained over the baseline for children's speech and adults' speech recognition under mismatched conditions.

1. Introduction

In recent years, development of speech recognition systems has enhanced the use of machines and other interactive multimedia systems in diverse areas [1]. Nowadays children have also become the potential users of these systems and, therefore, there is need for children's speech recognition. This will make their interaction with machines possible for various tasks like reading tutors, language learning by children, information retrieval, and entertainment applications [2–5]. Most speech recognition systems perform reasonably well for adult users but exhibit severe degradation in case of children users [6, 7]. Children's speech differs considerably from adults' speech in many important aspects and characteristics. The various acoustic and linguistic differences include differences in the pitch, the formant frequencies, the average phone duration, the speaking rate, the glottal parameters, pronunciation, and grammar. Children have a greater range of values with different means and variances for these parameters than adults due to anatomical and physiological changes occurring during a child's growth [8], thus resulting in a high inter- and intraspeaker acoustic variability. These differences together cause the deterioration in the recognition performance of children's speech on adults' speech trained models and vice versa [9, 10].

Children have nonlinearly increasing formants located at high values [11–13]. Also, they have high pitch frequency values causing large spacing between the harmonics [11–13]. These high frequency values of formants and pitch are attributed to their inherently short vocal tract and vocal fold lengths, respectively. For instance, five-year-old children have been reported to have 50% higher value of formant frequencies than adult males [8]. The higher formants of children fall outside the transmission bandwidth of telephone channel resulting in the loss of the spectral information in case of narrowband speech recognition. In comparison to the presence of 3-4 formants of an adult in 0.3–3.2 kHz bandwidth range, children have only 2-3 formants present [14]. The phoneme durations and the average sentence durations have also been observed to be

nearly 10% longer than those of adults [8, 11, 13, 15], which in turn reduce their speaking rate [7, 8]. The physiological differences among the speakers cause differences in the glottal parameters and thus the source spectrum [16]. For instance, the open quotient (OQ) mainly affects the levels of the lower part of the source spectrum so that a large OQ typically means a higher level of the lowest few harmonics. The return quotient (RQ) affects the steepness of the source spectrum, a large RQ corresponds to greater attenuation of the higher frequencies. These glottal parameters like the open quotient (OQ), the return quotient (RQ), and the speed quotient (SQ) have also been observed to be different for speech corresponding to children and adult speakers [17, 18]. Children exhibit less precision in the control of their articulators especially at the age of 5-6 years rendering to various pronunciation problems like disfluencies, false-starts, and extraneous speech [7]. Their vocabulary is smaller than that of adults and sometimes also contains some spurious words which are not found in the case of adults. It has been reported that children of the age of 5 years have about 60% of vowel classification accuracy against that of 90% of the adults [8].

Various methodologies and research issues have been investigated for improving children's speech recognition performance on adults' speech trained models. The foremost includes the vocal tract length normalization (VTLN) [7, 13]. It diminishes the effect of varying vocal tract length among different speakers by warping the frequency axis of the speech power spectrum during signal analysis [19]. Various forms of speaker adaptation techniques like maximum a posteriori (MAP) and maximum likelihood linear regression (MLLR) [13], speaker adaptive training (SAT) [20, 21], constrained MLLR speaker normalization (CMLSN) [22], and their combinations [13] have also been tried so as to reduce the mismatch of children's speech with adults' speech trained models. SAT performs speaker-specific transformations to compensate for the interspeaker acoustic variations in the training set [20]. It involves MLLR adaptation of the means of output distributions of continuous density hidden Markov models (HMMs). CMLSN method transforms the acoustic observation vectors by means of speaker-specific affine transformations obtained through constrained MLLR [13]. In order to cope with the age-dependent variability, age-specific modeling of recognizers has also been tried [3, 9, 23]. However, training age-specific speech models requires huge amount of data from the target age speakers, thus making the method costlier. To incorporate the linguistic mismatches between children's and adults' speech, language modeling [24, 25] and pronunciation modeling [26] have also been explored.

In contrast to various feature and model domain techniques, recent few studies have reported explicit normalization of various differences in the signal domain. A voice transformation technique which normalizes the children's speech signal before being fed to the adults' speech trained recognizer has been explored in [27]. It modifies the speech signal by transforming its pitch using the time-domain pitch-synchronous overlap-add (TD-PSOLA) method and obtaining VTLN by linear compression of the spectral envelope of each window. The use of the phase vocoder algorithm has also been demonstrated for achieving the same transformation. In [28], the speaking rate normalization in combination with VTLN has also been explored to achieve a better performance for children's speech on adults' speech trained recognizer.

Motivated by the studies done in [27, 28], this paper explores the independent effect of all of the acoustic sources of mismatch between adults' and children's speech reported in literature, that is, the pitch, the speaking rate, the formant frequencies, and the glottal parameters: OQ, RQ, and SQ on the recognition performance on a linguistically neutral task. Among these different acoustic sources of mismatch, the independent effects of the pitch and the glottal parameters on ASR have not been reported so far. The study is done on a limited vocabulary task (i.e., digit recognition) where the linguistic differences would be minimal.

The rest of the paper is organized as follows. Section 2 describes the technique used for transformation of different acoustic parameters of speech signals. Section 3 presents the details about the speech corpus and the experimental setup. Section 4 studies the degree of variation in various acoustic correlates between the adults' and the children's speech data used in this work. Section 5 discusses the results of the recognition experiments and the paper concludes in Section 6.

2. Transformation Procedures

In this work, the pitch, the signal duration (for modifying the speaking rate), and the glottal parameters, namely, the OQ, the RQ, and the SQ of the speech signals, are modified using a recently proposed pitch-synchronous time-scaling (PSTS) method [29]. The PSTS method is reported to provide faithful transformations over a wide range of transformation factors for the abovesaid parameters.

For addressing the mismatch in the formant frequencies between adults' and children's speech, the commonly used frequency warping is employed. For warping the frequency axis of the utterances during computation of the mel frequency cepstral coefficients (MFCCs) feature, the piece-wise linear frequency warping of filterbank, as supported in the hidden Markov toolkit (HTK) [30], has been used. In the following subsections, we describe the use of PSTS method for transforming the average pitch, the signal duration, and the glottal parameters (OQ, RQ, and SQ) of the speech signals.

2.1. PSTS Method. The PSTS method involves pitch-synchronous-time scaling of the linear prediction (LP) residual waveform of the speech signal. By time-scaling the short-time signals, the overlapping interval can be changed maintaining the energy balance of the modified signal. Since the LP residual signal approximates the derivative of the excitation signal, the time scaling operation also helps in preserving various important parameters of the glottal waveform like the OQ, the RQ, and the SQ. Additionally, it also overcomes the problem of energy fluctuations at large pitch modification factors which have been observed

in case of pitch transformation using the pitch-synchronous overlap-add-based approaches [29].

For doing the pitch-synchronous LP analysis, the pitch marks in the voiced regions are computed by glottal closure instants (GCIs) estimation algorithm, and in the unvoiced regions, the pitch marks are kept 5 ms equispaced. A 10th order pitch-synchronous LP analysis of the speech signal is performed using a 20 ms Hanning window centered on each pitch mark estimate. The residual signal is obtained by inverse filtering of the speech signal by a time-varying all-zeros filter defined by the linear prediction coefficients (LPCs) associated with each pitch mark. The analysis short-time signals, $\text{res}_i(n)$, are obtained by shifting the LP residual signal, $\text{res}(n)$, to begin in the previous analysis pitch mark, $p_a(i-1)$, and then multiplying it with a rectangular window, $\text{rec}(n)$, of length equal to the analysis pitch period, $P_a(i) = p_a(i) - p_a(i-1)$:

$$\text{res}_i(n) = \text{res}\big(n + p_a(i-1)\big)\text{rec}(n), \tag{1}$$

where

$$\text{rec}(n) = \begin{cases} 1, & 0 \le n < P_a(i), \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

2.1.1. Pitch and Signal Duration Transformation. The pitch marks and the LP residual signal are computed as described in Section 2.1. The modified pitch mark locations are then computed in accordance to the desired pitch and signal duration (for speaking rate) modification. The shift between successive synthesis pitch marks is equal to the desired pitch period $P_s(j) = p_s(j) - p_s(j-1)$. The short-time signals $\text{res}_j(n)$ are computed by mapping the synthesis pitch marks $p_s(j)$ on the estimated analysis pitch marks $p_a(i)$. Each short-time signal $\text{res}_j(n)$ is time scaled by a factor $N = P_s(j)/P_a(i)$ resulting in the modified short-time signal $x_j(n)$ where $P_s(j)$ is the desired synthesis pitch period.

The pitch and duration transformations can lead to either removal or replication of the analysis short-time signals according to the modification factor. To avoid various phase and frequency discontinuities in the energy envelope and achieve smooth spectral transitions, the nonadjacent lth and rth short-time analysis signals are first time scaled to the desired synthesis pitch period $P_s(j)$ to get $x_l(n)$ and $x_r(n)$. Then, $x_l(n)$ and $x_r(n)$ are weighted and added to obtain the resultant modified jth short-time synthesis signal $y_j(n)$:

$$y_j(n) = h(n)x_l(n) + h\big(P_s(j) - n\big)x_r(n), \tag{3}$$

where $h(n)$ is a decaying weighting window of size equal to $P_s(j)$, such that $h(n) + h(P_s(j) - n) = 1$. The right half of the Hanning window satisfies this condition.

Finally, the complete synthesis LP residual signal is obtained by the pitch-synchronous sum of all of the synthesis short-time signals using (4). The modified speech signal is synthesized by passing the modified LP residual through a time-varying all-zeros filter defined by the LPC mapped to the synthesis pitch marks:

$$y(n) = \sum_j \Big(y_j\big(n - p_s(j) + P_s(j)\big)\Big). \tag{4}$$

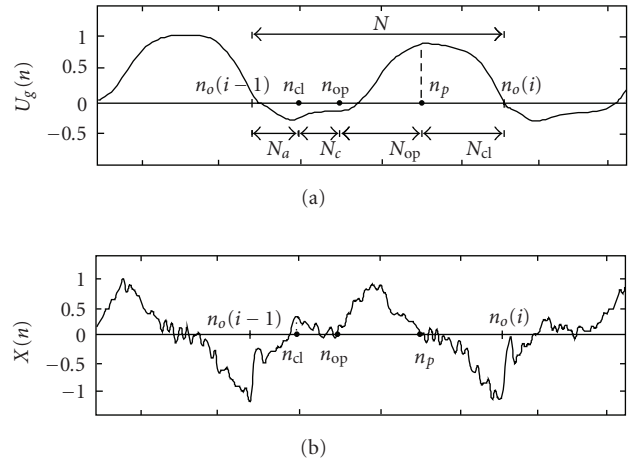

Figure 1: Representation of the extracted time instants and the glottal cycle phases in (a) the glottal flow waveform and (b) its time-derivative (i.e., the LP residual signal). The figure is adapted from [29].

For pitch transformation factors $N > 1$, the spectra of the speech signal get compressed, giving rise to an "energy hole" at the higher frequencies. Since in our experiments children's speech is transformed to adults' speech, this problem gets even more enhanced. To overcome this problem, a high-frequency regeneration method based on time-scaling the open phase of the glottal source waveform has been used so as to reduce the OQ which in turn boosts the energy at the high-frequency region of the source spectrum to fill the energy hole. For details of the high-frequency regeneration method refer to [31].

2.1.2. Glottal Parameter Transformation. The pitch marks and the LP residual signal are computed as described in Section 2.1. Corresponding to each pitch cycle a short-time analysis frame is determined using (1). The following time instants are then estimated for each of the voiced short-time analysis frames: the glottal closure instant (n_{cl}), the glottal opening instant (n_{op}), and the instant of maximum of the glottal flow (n_p).

In order to transform the glottal flow parameters, time scale transformations are done over the segments corresponding to the glottal flow phases in each of the short-time analysis frames. The segments corresponding to each of the glottal cycle phases are computed from the extracted time instants using following relations:

Return Phase:

$$N_a = n_{\text{cl}}, \tag{5}$$

Peak Flow Duration:

$$N_e = N - n_{\text{op}}, \tag{6}$$

Closed Phase:

$$N_c = N - N_a - N_e, \tag{7}$$

Table 1: Age groupwise breakup of the children's speech data.

	Age group (Yrs.)				
	6-7	8-9	10-11	12-13	14-15
No. of Speakers	8	31	42	17	3
(Boys/Girls)	(5/3)	(12/19)	(27/15)	(5/12)	(1/2)
No. of Utterances	615	2386	3231	1309	231

Opening Phase:

$$N_{\text{op}} = n_p - n_{\text{op}}, \tag{8}$$

Closing Phase:

$$N_{\text{cl}} = N - n_p. \tag{9}$$

A typical illustration of the various extracted time instants and the glottal cycle phases is given in Figure 1.

The open quotient is related to the duration of the open phase and can be expressed as

$$\text{OQ} = \frac{(N_a + N_e)}{N}. \tag{10}$$

To increase OQ, both the return phase duration and the peak flow duration must be increased. To decrease OQ, both of the durations must be shortened. Thus, the time scale factor is equal to the required modification factor for OQ. Due to time scale transformation it is necessary to adjust the duration of the closed phase to preserve the pitch period of the glottal waveform as described in [29].

The return quotient is related with the duration of the closing phase and determines the cutoff frequency of the spectral tilt. It is computed as.

$$\text{RQ} = \frac{N_a}{N}. \tag{11}$$

The return quotient can be increased or decreased by a time scale expansion or compression of the return phase. To maintain the pitch period and the open quotient, the peak flow duration is also time scaled by an adequate factor.

The speed quotient is related to the asymmetry coefficient and accounts for variations in the shape of the segment corresponding to the open phase of the glottal flow. It can be expressed as

$$\text{SQ} = \frac{N_{\text{op}}}{N_{\text{cl}}}. \tag{12}$$

The speed quotient can be increased with a time scale expansion of the opening phase and a time scale compression of the closing phase so that the peak flow duration $N_e = N_{\text{op}} + N_{\text{cl}}$ remains constant. SQ can be decreased by the opposite transformation.

Finally, the complete synthesis LP residual signal and the modified synthesis speech signal are computed as described in Section 2.1.1. The sample speech files with the average pitch, the average utterance duration, and the average values of the glottal parameters (OQ, RQ, and SQ) modified by different factors are available at "http://www.iitg.ac.in/ece/emstlab/psts.htm" for assessing the quality of the various transformations.

3. Speech Corpus and Experimental Setup

In this work, to assess the impact of various acoustic parameters of speech signal on the speech recognition performance the automatic speech recognition (ASR) systems are developed using the TIDIGITS [32]. The TIDIGITS database contains 11.4 hours of speech data from 326 speakers (111 men, 114 women, 50 boys, and 51 girls) uttering one to seven digits long strings consisting of eleven different digits (0–9 and "OH"). The range of the age of the adult speakers is from 17 years to 70 years while the children speakers belong to an age group of 6 years to 15 years. All speech data is downsampled from 20 kHz at 8 kHz for use in this work. The age groupwise details of the complete children's speech data are given in Table 1.

For experiments done on adults' speech trained recognizer, the adults' speech training set referred to as "TR1", the adults' speech test set referred to as "AD", and the children's speech test set referred to as "CH1" have been derived from the TIDIGITS corpus. "TR1" comprises of the adults' speech data containing a total of 11,016 utterances, or 35,566 digits, from 90 male and 107 female speakers. "AD" comprises of the adults' speech data containing a total of 3,303 utterances, or 10,813 digits, from 29 male and 52 female speakers. "CH1" comprises of whole children's speech data containing a total of 7,772 utterances, or 25,525 digits, available from 50 boys and 51 girls.

For experiments done on children's speech trained recognizer, the children's speech training set referred to as "TR2", the children's speech test set referred to as "CH2", and the adults' speech test set referred to as "AD" have been derived from the TIDIGITS corpus. "TR2" and the "CH2" datasets have been derived by splitting the "CH1" dataset. "TR2" comprises of the children's speech data containing a total of 4,481 utterances, or 14,725 digits, from 31 boys and 33 girls. "CH2" comprises of only the children's speech data containing a total of 3,291 utterances, or 10,800 digits, available from 22 boys and 27 girls. It is to note that all of these datasets are disjoint from each other in terms of the speech utterances but not the speakers. The details of all of the training and test speech sets used in the following experiments are summarized in Table 2.

Throughout this paper, word error rate (WER) is used to evaluate the performance of various techniques. The error rate is computed as follows:

$$\%\text{WER} = \frac{\text{Sub} + \text{Del} + \text{Ins}}{\text{Total No. of Words}} \times 100, \tag{13}$$

where "Sub" is the number of substitutions, "Del" is the number of deletions, and "Ins" is the number of insertions.

Table 2: Details of all of the training and test speech sets.

Dataset		No. of Utterances	No. of Speakers			
			Adults		Children	
			Men	Women	Boys	Girls
Adults	Train "TR1"	11,016	90	107	—	—
	Test "AD"	3,303	29	52	—	—
Children	Test "CH1"	7,772	—	—	50	51
	Train "TR2"	4,481	—	—	31	33
	Test "CH2"	3,291	—	—	22	27

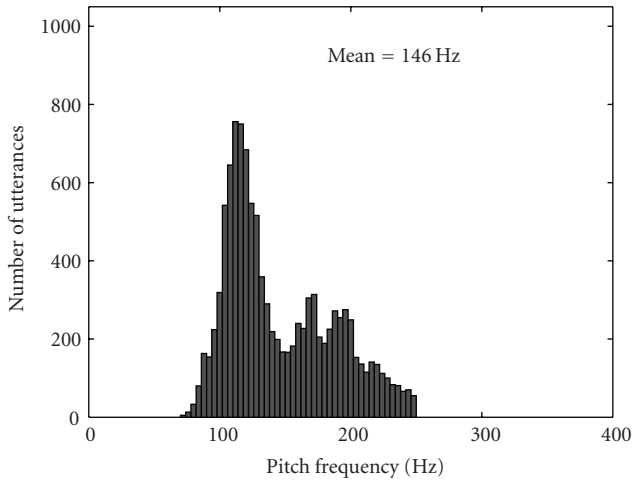

Figure 2: Distribution of the average pitch of the original signals of the adult training set "TR1".

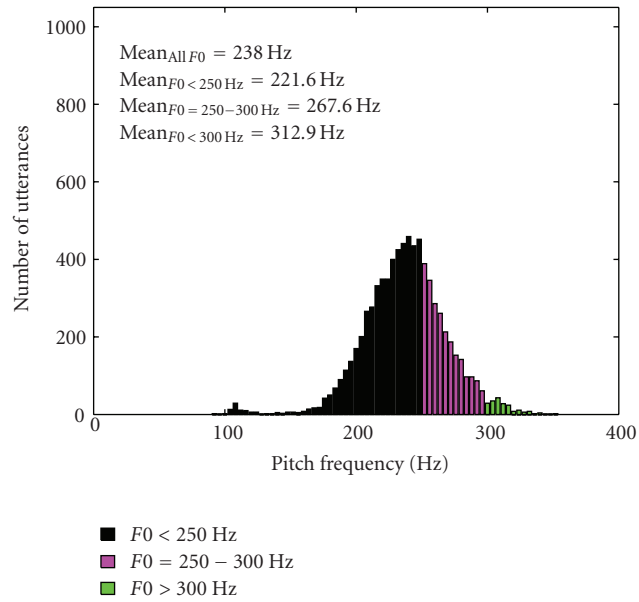

Figure 3: Distribution of the average pitch of the original signals of the children test set "CH1". Three broad pitch groups have been marked with three different colors for studying their distribution after pitch-normalization.

The connected digit recognizer used in this work has been developed using the HTK toolkit [30]. The 11 digits (0–9 and "OH") are modeled as whole word left-to-right hidden Markov model (HMM). Each word model has 16 states with simple left to right paths and no skip paths over the states. The observation densities are mixtures of five multivariate Gaussian distributions with diagonal covariance matrices. The silence is explicitly modeled using three-state HMM model having six Gaussian mixtures per state. A single-state short-pause model tied to the middle state of the silence model is also used. A 21-channel filterbank is used for 13-dimensional MFCC (C_0 to C_{12}) feature computation. In addition to the base features, their first- and second-order derivatives are also appended making the final feature dimension as 39. Cepstral mean subtraction is also applied to all features. The speech is preemphasized using a factor of 0.97 and for analysis a Hamming window of length of 25 ms and the frame rate of 100 Hz is used.

4. Acoustic Analysis of the Speech Database

In this section, we quantify the degree of mismatch in various acoustic correlates of the adults' and the children's speech data used for the recognition experiments in this work. This is done in order to hypothesize the relative effect of normalization of each of these acoustic correlates on the ASR performance under mismatched conditions. The various acoustic correlates that have been analyzed include the pitch, the speaking rate, and the glottal parameters (OQ, RQ, and SQ).

4.1. Pitch. The difference in the average pitch values of the children's and the adults' speech data used in this work can be understood by observing the distribution of the average pitch of the signals of the adults training set "TR1" and the children test set "CH1" as shown in Figures 2 and 3, respectively. It is noted that the mean of the pitch distribution of the children test set "CH1" is nearly of the order of 1.6 to that of the adults training set "TR1". Thus, as expected, children have significantly higher pitch values than adults.

In [33], we have explored the effect of pitch variations on MFCC feature. Our study reveals that as the pitch of the signals increases some pitch-dependent distortions appear in the spectrum corresponding to MFCC feature particularly at lower frequencies up to 1 kHz. These distortions can be seen in the smooth spectrum of the stable voiced portion

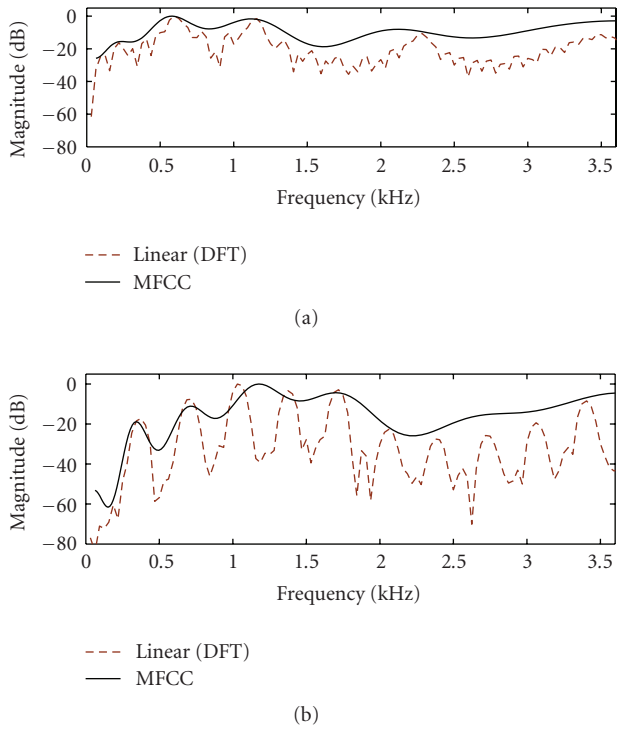

Figure 4: Plots of smooth spectra corresponding to MFCC feature along with linear (DFT) spectra of a stable voiced frame extracted from the digit "OH" signals having average pitch value of (a) 85 Hz and (b) 310 Hz.

extracted from the digit "OH" signal having average pitch value 310 Hz when compared with that of the 85 Hz average pitch signal as shown in Figure 4. The smooth spectrum corresponding to MFCC is derived by computing a 128-point inverse discrete cosine transform of 13-dimensional MFCC feature (C_0–C_{12}). The cause of these distortions in the spectral envelope is attributed to the insufficient smoothing of the pitch harmonics by the filterbank particularly at low frequencies as the width of the lower-order filters is around 100 Hz only.

Therefore, on noting the high degree of variation in the average pitch values of the children's and the adults' speech data and the effect of high pitch value on the smooth spectrum corresponding to MFCC feature, it is hypothesized that pitch-normalization would significantly improve the ASR performance for children's speech due to reduction in the pitch-dependent distortions observed in the spectral envelope.

4.2. Speaking Rate. The difference in the average speaking rate of the children's and the adults' speech data used in this work can be understood by observing the distribution of the average speaking rate of the adults training set "TR1" and the children test set "CH1" as shown in Figures 5 and 6, respectively. It is noted that the mean of the speaking rate distribution of the adults training set "TR1" is 1.2 times that of the children test set "CH1". Thus, as expected, children

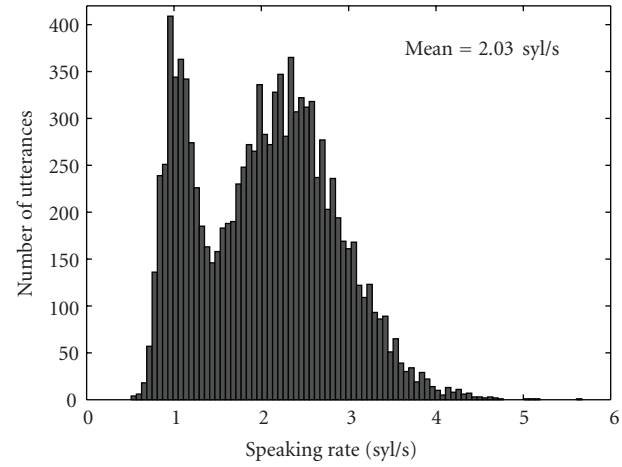

Figure 5: Distribution of the speaking rate of the original signals of the adult training set "TR1".

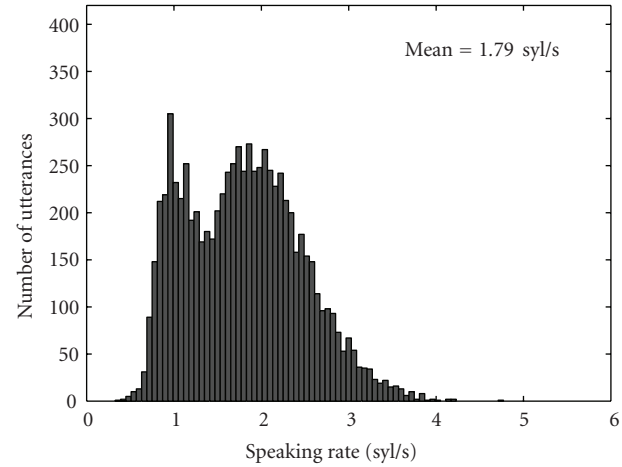

Figure 6: Distribution of the speaking rate of the original signals of the children test set "CH1".

have longer sentence duration, and thus lower speaking rate than adults.

It has been reported in perceptual studies that the variation in the speaking rate affects the acoustic patterns of speech by restructuring the relationship between the acoustic cues and the phonetic categories [34]. It has also been shown that when speaking rate increases the duration of the vowels is affected the most [35]. So, the models, trained with fast and slow speaking rate speech data, would have different transition probabilities. The models trained with speech data having slow speaking rate would have greater self-loop transition probabilities than the models trained with speech data having fast speaking rate. In order to verify this, we compared the self-loop transition probability of each of the 16 states of the single digit "OH" model corresponding to both the speech data with slow speaking rate (children test set "CH1") and the speech data with fast speaking rate (adults training set "TR1") as shown in Figure 7. It is noted that, as

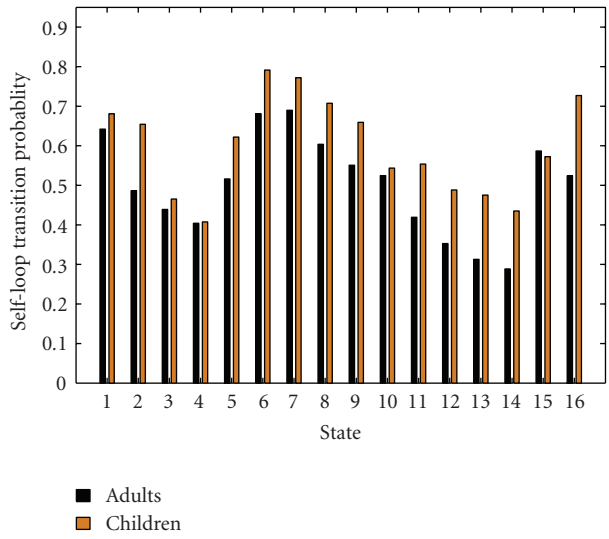

Figure 7: Statewise self-loop transition probabilities of the digit "OH" models corresponding to the adults dataset "TR1" and the children dataset "CH1".

expected, the children's speech trained model has greater self-loop transition probability across all states in comparison to that of the adults' speech trained model due to longer sentence/phone durations.

Thus, the state transition probabilities of models trained on speech data with fast speaking rate (adults' speech) would adversely affect the ASR performance for speech data with slow speaking rate (children's speech) [36–38]. So, in order to recognize a particular acoustic property as the intended phonetic segment, it is required to normalize the speaking rate differences.

4.3. Glottal Parameters. The difference in the average value of the three glottal parameters of the children's and the adults' speech data used in this work can be understood by observing the distribution of the average values of the OQ, RQ, and SQ for both the adults training set "TR1" and the children test set "CH1" as shown in Figures 8, 9, and 10, respectively. It is noted that the mean values of all the three glottal parameters for the adults training set "TR1" are smaller than those for the children test set "CH1". Thus, as expected, children have more breathiness in their speech than adults.

In order to hypothesize the effect of normalization of each of these three glottal parameters on the ASR performance of the children test set "CH1" under mismatched condition, the smooth spectra corresponding to MFCC obtained from speech signals with original glottal parameter values and with transformed glottal parameter values are compared. The smooth spectra for the stable voiced portions extracted from the single digit "OH" signals with original and transformed values of OQ, RQ, and SQ are shown in Figure 11. It is noted that transformation of each of the three glottal parameters gives rise to some changes in the smooth spectra corresponding to MFCC but not in any systematic manner. Thus, it is hypothesized that the normalization of these glottal parameters for the children's speech signals may not significantly affect their ASR performance under mismatched conditions.

5. Experimental Results and Discussion

This section describes our experiments to study the effect of various acoustic correlates in addressing the mismatch between adults' and children's speech on their recognition performance under mismatched conditions. This study is first explored in detail for addressing the recognition of children's speech on adults' speech trained models. Following it we have also shown the results of a similar study on vice-versa condition.

5.1. Children's Speech Recognition on Adults' Speech Trained Models. The adults' speech trained models used in this study have been developed using the adults training set "TR1" derived from the TIDIGITS corpus. The recognition performances for the adult test set "AD" and the children test set "CH1" are 0.43% and 11.37%, respectively.

For normalizing the mismatch of the children's speech signals with respect to the adults' speech trained acoustic models, various acoustic correlates need to be modified appropriately. To determine the optimal value to which each of the acoustic correlates is to be transformed, a maximum-likelihood- (ML-) based grid search is used. For instance, the optimal value of an acoustic correlate, say $\hat{\alpha}$, given its various transformed values within the valid range, is estimated as

$$\hat{\alpha} = \arg\max_{\alpha} Pr\left(X_i^{\alpha} \lambda, W_i\right), \tag{14}$$

where X_i^{α} is the feature corresponding to a particular value α of an acoustic correlate for the ith utterance, λ is the speech recognition model, and W_i is the transcription of the ith utterance. The W_i is determined by doing an initial recognition pass using original feature (i.e., with no transformation).

The appropriate values for transformation of the average pitch frequency, the signal duration (and thus the speaking rate), the glottal source parameters (OQ, RQ, SQ), and the formant frequencies are obtained using the above procedure. In this work, the average pitch frequency, the signal duration, and the glottal source parameters are transformed explicitly in the signal domain prior to feature computation whereas the formant frequencies are modified in the feature domain. The average pitch of a speech signal is estimated using the ESPS tool available in the Wavesurfer software package [39]. The average speaking rate of the signals is measured as the number of syllables per second computed as the ratio of the number of syllables in an utterance to the total length of the utterance. Each of the 11 digits constituting the training and test set utterances used in this work comprises of only 1 syllable except for the digits "zero" and "seven" which contain 2 syllables. In the following, we describe in detail the experimental conditions and the results obtained by the normalization of each of these acoustic correlates independently as well as in combinations.

5.1.1. Pitch. For pitch-normalization of the children test set "CH1", the signals are transformed to seven different pitch

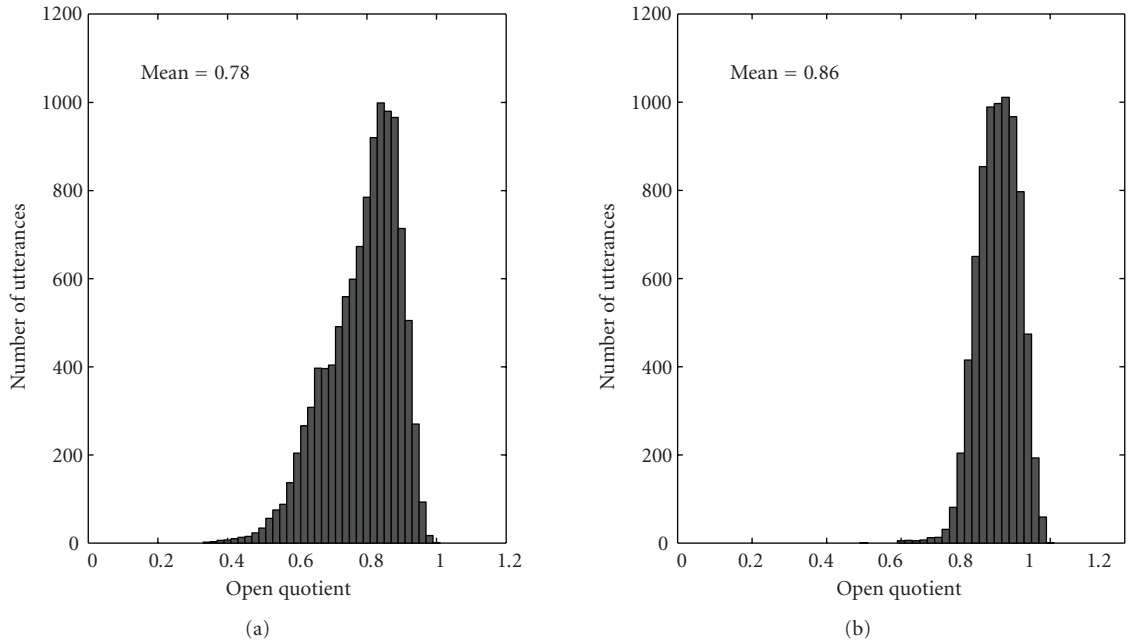

Figure 8: Distribution of the average OQ values of the original signals of the (a) adult training set "TR1" and (b) children test set "CH1".

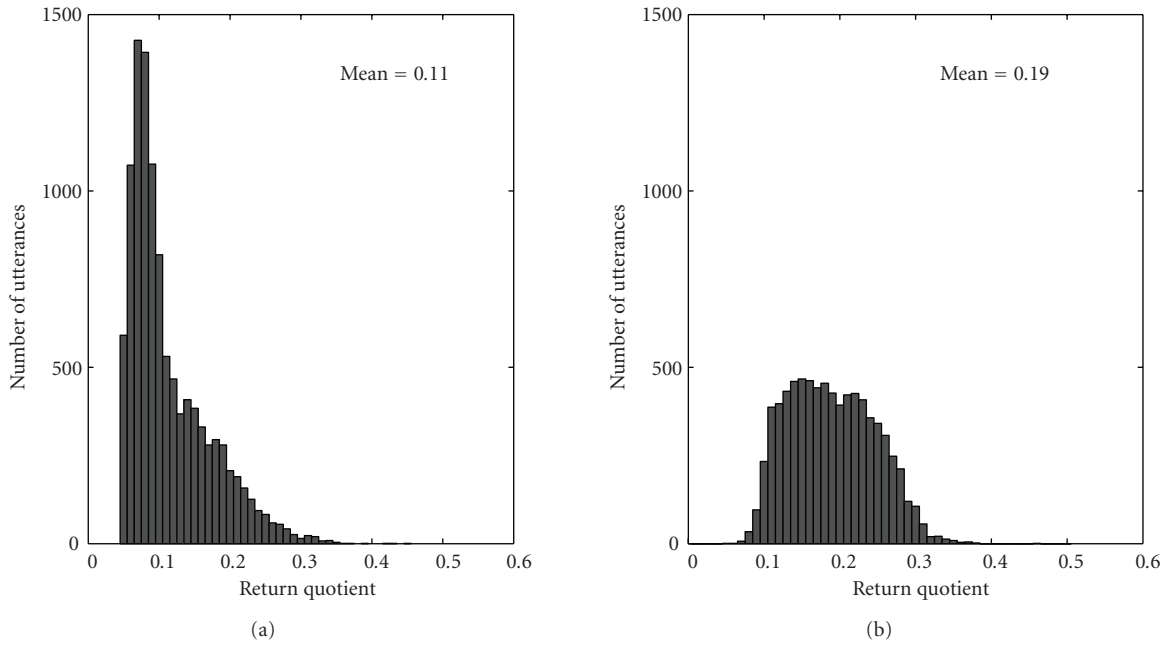

Figure 9: Distribution of the average RQ values of the original signals of the (a) adult training set "TR1" (b) children test set "CH1".

values ranging from 70 Hz to 250 Hz with a step size of 30 Hz. Such pitch range has been chosen based on the pitch distribution of the training data as shown in Figure 2.

The recognition performances of the children test set "CH1" with and without pitch-normalization are given in Table 3. It is noted that pitch-normalization results in 15% relative improvement over the baseline performance. On observing the pitch groupwise performances, also given in Table 3, for test signals having average pitch value before transformation in the range of <250 Hz, 250–300 Hz, and >300 Hz, a relative improvement of about 8%, 19%, and 23% is obtained after pitch-normalization, respectively. Thus, consistent improvements are noted for the different pitch groups; that is, higher pitch groups have greater improvements. The improvements obtained with the pitch-normalization can be further understood by observing the

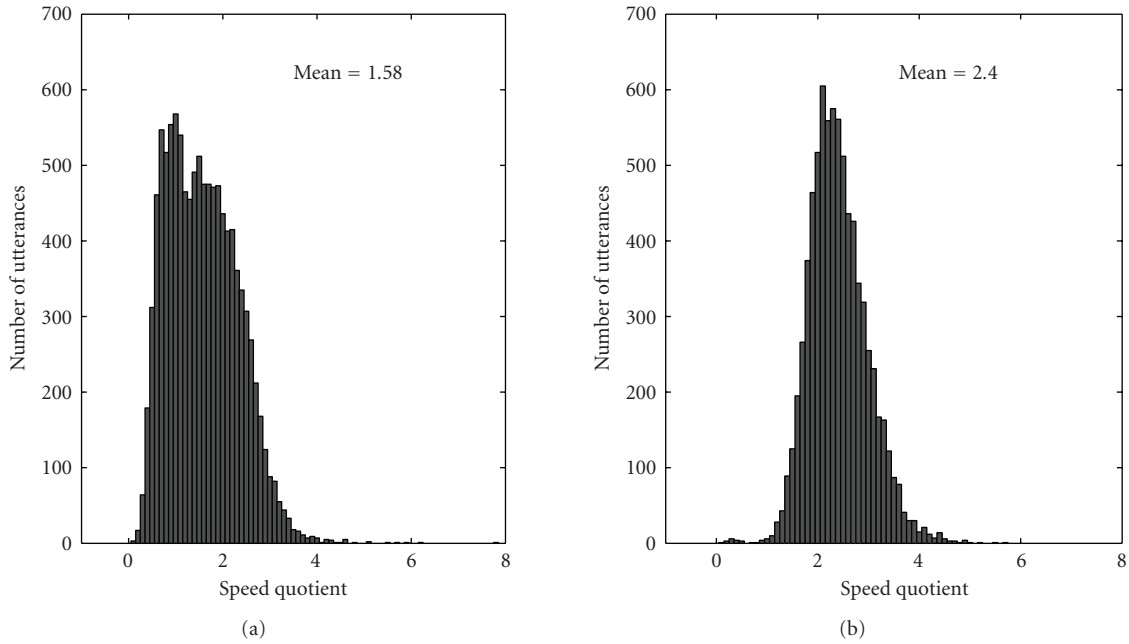

Figure 10: Distribution of the average SQ values of the original signals of the (a) adult training set "TR1" and (b) children test set "CH1".

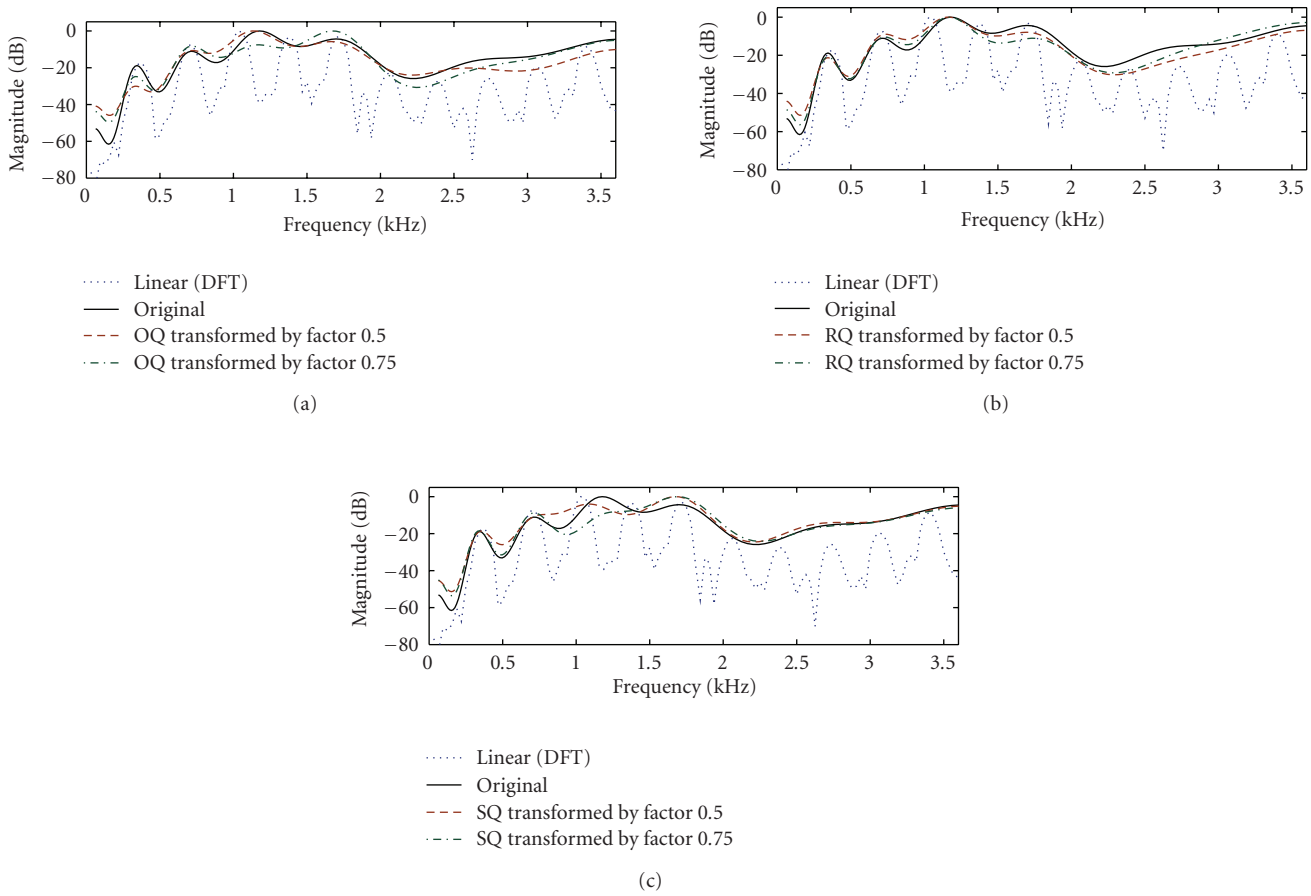

Figure 11: Plots of the original linear (DFT) spectrum along with the smooth spectra corresponding to MFCC feature of a digit "OH" signal with original and transformed values of (a) OQ, (b) RQ, and (c) SQ.

Table 3: Performances of the children test set "CH1" (with breakup for different pitch groups based on original average pitch values) with and without pitch-normalization. The quantity in bracket shows the number of utterances in that group. The 95% confidence interval for the performance is 0.39 (for the <250 Hz, 250–300 Hz, and >300 Hz pitch groups the confidence interval turns out to be 0.39, 0.79, and 3.37, resp.).

Condition	WER (%)			
	All F_o Values (7,772)	$F_o < 250$ Hz (5,224)	$F_o = 250$–300 Hz (2,346)	$F_o > 300$ Hz (202)
Baseline	11.37	6.54	17.47	39.03
Norm.	9.64	6.02	14.24	30.11

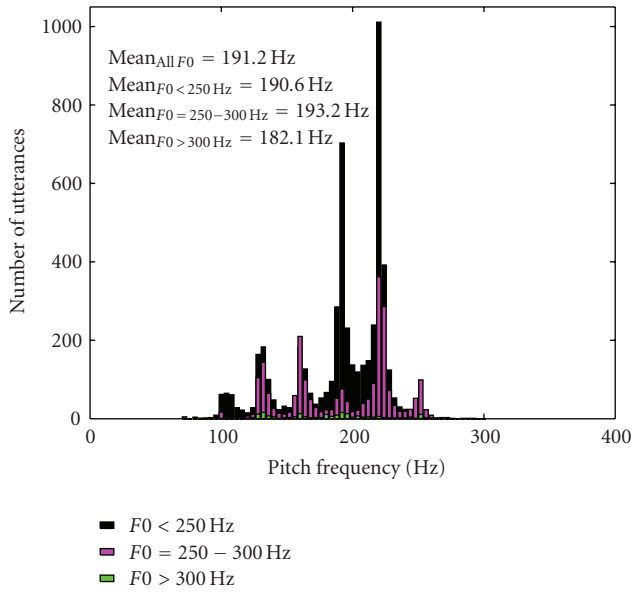

Figure 12: Distributions of the average pitch of the signals of the children test set "CH1" after ML-based pitch-normalization for all of the three pitch groups (as defined in Figure 3 based on their original pitch values) plotted separately.

distribution of the pitch of the adults training set "TR1" and the children test set "CH1" before and after pitch-normalization as shown in Figures 2, 3, and 12, respectively. It is noted that the mean of the pitch distribution of the children test set "CH1" has shifted towards that of the adults training set "TR1" due to ML-based pitch-normalization. Also, comparing the mean of the pitch distribution for all pitch group signals before and after pitch-normalization, it is noted that the shift in the mean of the pitch distribution is more for higher pitch groups.

These improvements in the recognition performance of children's speech with pitch-normalization could be attributed to the reduction of the earlier hypothesized pitch-dependent distortions that occur in the smooth spectral envelope corresponding to the MFCC feature due to insufficient smoothing of the pitch harmonics by the filterbank, particularly for high pitch signals. This can be further verified by comparing the smooth spectrum of a stable voiced portion extracted from a digit "OH" signal having original average pitch value of 310 Hz with that obtained after reducing the average pitch of the signal by a factor of 1.6 as shown in Figure 13. It is noted that, as hypothesized, after pitch reduction the pitch-dependent distortions observed in the spectral envelope at the lower frequencies (below 1 kHz) for the original 310 Hz pitch signal are significantly reduced.

5.1.2. Speaking Rate. For normalization of the speaking rate of the children test set "CH1" according to that of the adults' speech trained models, the duration of the signals is reduced by factors ranging from 0.6 to 1 with a step size of 0.05, thereby increasing the speaking rate of the signals by factors ranging from 1 to 1.65. The choice of such duration transformation factors is based on the distribution of the speaking rate of the signals belonging to the adults training set as shown in Figure 5.

The recognition performances of the children test set "CH1" with and without speaking rate normalization are given in Table 4. It is noted that speaking rate normalization results in a significant 9% relative improvement over the baseline performance. This improvement is consistent with the results reported in literature obtained with speaking rate normalization for ASR of signals of mismatched speaking rate [28, 36]. The distributions of the speaking rate of the adults training set "TR1" and the children test set "CH1" before and after speaking rate normalization are shown in Figures 5, 6, and 14, respectively. On comparing these distributions, it is noted that the mean speaking rate of the children test set "CH1" has been transformed towards that of the adults training set "TR1" after speaking rate normalization.

Further, on comparing the likelihood of the signals from the children test set "CH1" on the adults' speech trained models before and after speaking rate normalization, as shown in Figure 15, it is noted that the likelihood of all of the children's speech utterances has increased after normalization of their speaking rate. This verifies the reduction in the earlier hypothesized mismatch in the duration modeling of the children test set with respect to the adults' speech trained models.

5.1.3. Glottal Parameters. For normalizing the variations in the glottal parameters of the children test set "CH1" with respect to those of the adults training set "TR1" used for training the acoustic models using the ML-based approach, the OQ, RQ, and SQ of the signals are modified by factors ranging from 0.55 to 1, 0.35 to 1, and 0.45 to 1, each with a step size of 0.05, respectively. The choice of such transformation factors for OQ, RQ, and SQ modification is supported by the studies done in literature which report that children's speech has higher OQ, RQ, and SQ values than those of the adults' speech [17, 18, 40]. The recognition

Table 4: Performances of the children test set "CH1" with and without normalization of different acoustic correlates of speech. The 95% confidence interval for the performances is 0.39.

Condition	WER (%)
Baseline	11.37
Norm. (Speaking Rate)	10.31
Norm. (Open Quotient)	11.32
Norm. (Return Quotient)	11.28
Norm. (Speed Quotient)	11.01
Norm. (Formant Frequencies)	2.95

performances of the children test set "CH1" with and without normalization of the three glottal parameters are given in Table 4. As hypothesized earlier, none of the glottal parameters give any significant improvement over the baseline after normalization. Although the glottal parameters have been found to be of significance in case of one-to-one voice transformation in case of ASR, where the acoustic model is trained using data from a large number of speakers, there is enough variation in the glottal parameters within the training set itself leaving a very little mismatch due to differences in the glottal parameters between the training and the test data.

The age groupwise distributions of the ML-based transformation factors chosen for normalization of OQ, RQ, and SQ of the children's test speech signals with respect to the adults' speech trained models are shown in Figures 16, 17, and 18, respectively. It is noted that in ML search for the optimal transformation factor, for normalization of each of these glottal parameters, majority of the signals have opted for no transformation across all age groups for all of the three glottal parameters. Also, it is worth noting that all transformation factors have been chosen by the signals of all age groups in similar proportion. Thus, there seems to be very little correlation between the age and the glottal parameters (OQ, RQ, SQ).

5.1.4. Formant Frequencies. The variation in the formant frequencies among adults' and children's speech occurs due to differences in their vocal tract lengths, which is usually modeled as a constant scaling of the resonant peaks in the spectral domain. For normalizing the variations in the formant frequencies of the signals of the children test set "CH1", an ML search is performed among features warped by 13 equally spaced warping factors ranging from 0.88 to 1.12 with a step size of 0.02 for each signal. The recognition performances of the children test set "CH1" with and without VTLN are given in Table 4. It is noted that VTLN results in a 74% relative improvement over the baseline performance which is highly significant as compared to the previous parameters studied. The improvements obtained with VTLN can be further understood by observing the distribution of the ML-based warping factors chosen for VTLN of the original children's speech signals with respect to the adults' speech trained models as shown in Figure 19. It is noted that majority of warp factors for the children test set

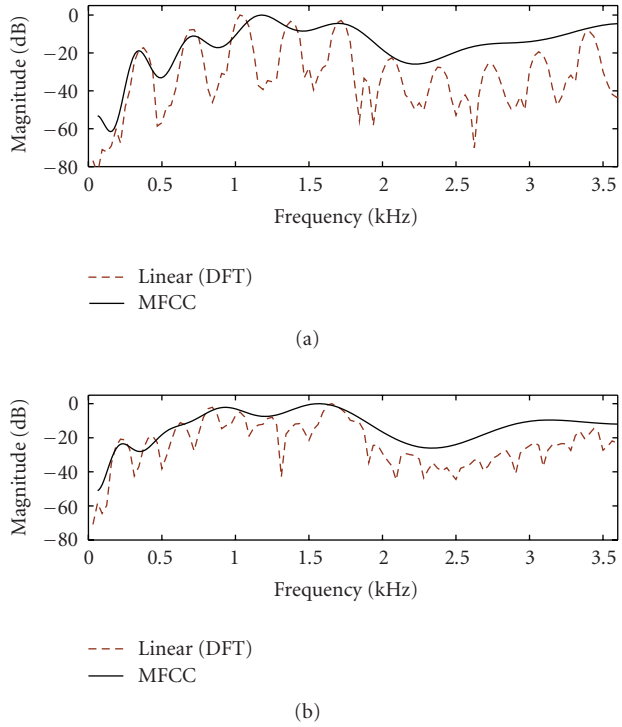

Figure 13: Plots of smooth spectra corresponding to MFCC feature along with linear (DFT) spectra of a stable voiced frame extracted from the digit "OH" signals having average pitch value of (a) 310 Hz (original), and (b) 190 Hz (after transformation of the average pitch of the signal from 310 Hz by factor of 1.6).

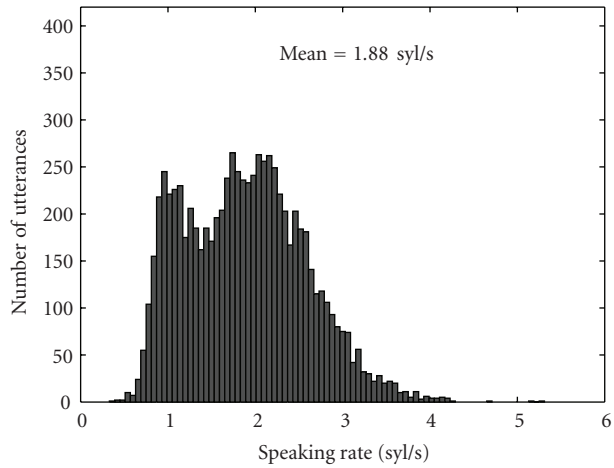

Figure 14: Distribution of the speaking rate of signals of the children test set "CH1" after ML-based rate normalization.

"CH1" are estimated as <1 (i.e., compression of the spectra), which is consistent with the fact that children have smaller vocal tract lengths than adults.

5.1.5. Combined Normalization of Acoustic Correlates. From the study done in the previous subsections analyzing the independent effect of each of the acoustic correlates of speech like the pitch, the speaking rate, the glottal parameters, and the formant frequencies on children's speech

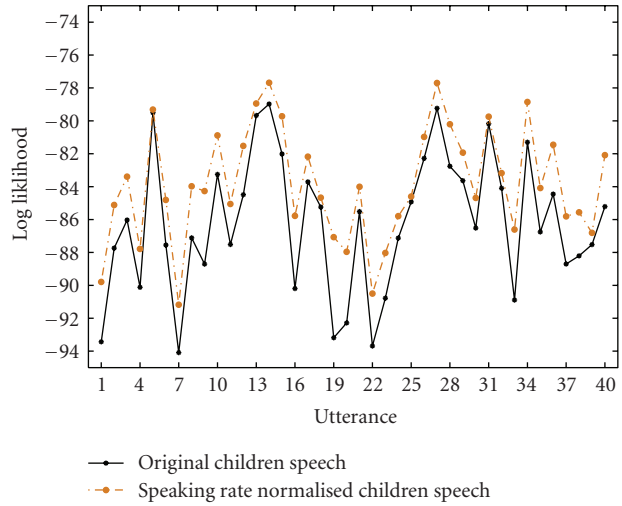

Figure 15: Log likelihood distribution of few utterances from the children's test set "CH1" before and after speaking rate normalization on models trained with adults training set "TR1".

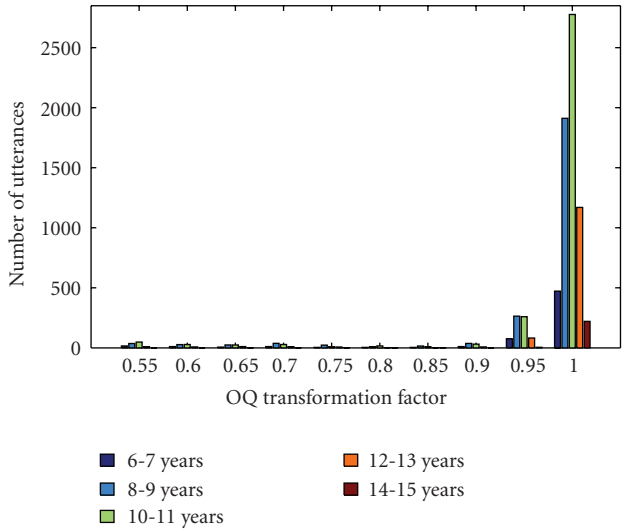

Figure 16: Age groupwise distribution of the optimal OQ transformation factors chosen for the signals of the children test set "CH1".

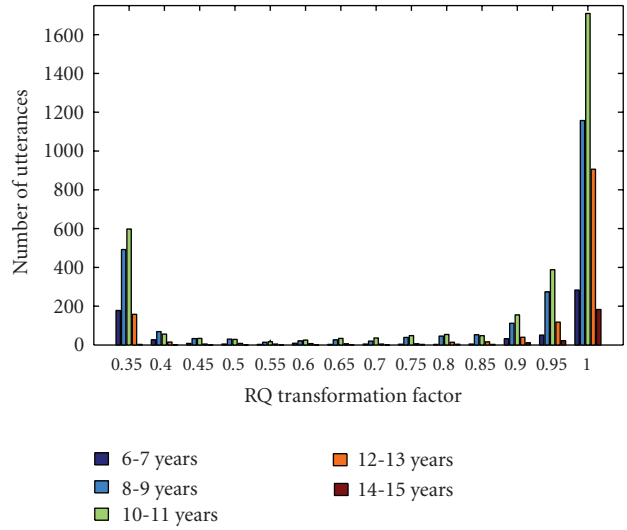

Figure 17: Age groupwise distribution of the optimal RQ transformation factors chosen for the signals of the children test set "CH1".

Table 5: Performances of the children test set "CH1" with and without normalization of different acoustic correlates of speech in various combinations. The 95% confidence interval for the performances is 0.39.

Condition	WER (%)
Baseline	11.37
Norm. (Speaking Rate + Pitch)	9.48
Norm. (Speaking Rate + Formant Freq.)	2.37
Norm. (Pitch + Formant Freq.)	3.70
Norm. (Speaking Rate + Pitch + Formant Freq.)	3.55
Norm. (Rate + Formant Freq.) (Using "Back Off" Procedure)	2.28
Norm. (Pitch + Formant Freq.) (Using "Back Off" Procedure)	2.46
Norm. (Speaking Rate + Pitch + Formant Freq.) (Using "Back Off" Procedure)	2.25

recognition performance, it is noted that the significant improvements in the recognition performance are obtained with the normalization of the pitch, the speaking rate, and the formant frequencies only. In this subsection, we study the effect of combined normalization of only these three acoustic correlates of speech on children's speech recognition performance. The combined normalization of the acoustic correlates of speech has been done in sequential manner, that is, for obtaining both the speaking rate and the pitch-normalized speech signals; first the speaking rate of the speech signal is normalized followed by its pitch-normalization. As mentioned earlier, VTLN is performed in the feature domain whereas the speaking rate and the pitch-normalization are done in the signal domain. Thus, to incorporate VTLN in combination with the speaking rate and the pitch-normalization, the signal is first speaking rate and/or pitch-normalized followed by VTLN. Between the speaking rate and the pitch-normalization, we have first normalized the speaking rate since the speaking rate transformation may result in slight modification of the pitch of the signals.

The recognition performances of the children test set "CH1" with normalization of various combinations of the said three acoustic correlates of speech are given in Table 5. It is noted that, on performing both the speaking rate and the pitch-normalization of children's speech, about 16% relative improvement is obtained over the baseline in children's speech recognition performance. On combining the speaking rate normalization along with VTLN, about 19% relative improvement is obtained over the recognition performance obtained with only VTLN of the original children's speech signals. This suggests that speaking rate normalization is additive to the improvement obtained by

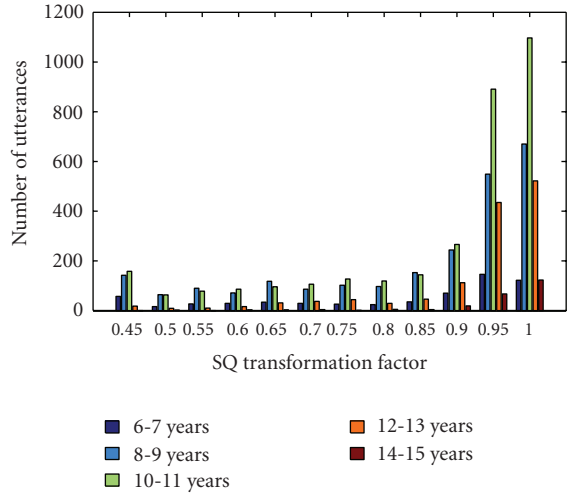

Figure 18: Age groupwise distribution of the optimal SQ transformation factors chosen for the signals of the children test set "CH1".

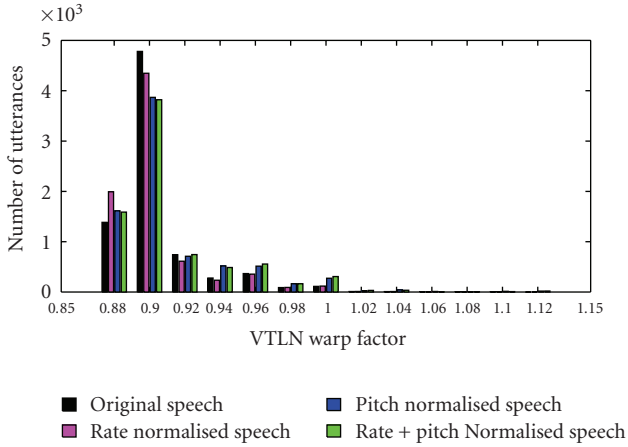

Figure 19: Distribution of the warping factors chosen for VTLN of the original, the speaking rate normalized, the pitch-normalized, and the combined speaking rate and pitch-normalized signals of the children test set "CH1".

VTLN of the children's speech signals. On the other hand, the combination of the pitch-normalization with VTLN and that of both the speaking rate and the pitch-normalization with VTLN show degraded performances as compared to that obtained with only VTLN of the original children's speech signals. Since the combined normalization of speaking rate and VTL shows a significant improvement in the recognition performance, this degradation due to the inclusion of the pitch-normalization is further investigated.

Figure 19 shows the distribution of the warp factors estimated for VTLN of the speaking rate normalized, the pitch-normalized and the combined speaking rate and pitch-normalized signals of the children test set "CH1". It is noted that after speaking rate normalization a larger number of signals have chosen optimal warp factors of 0.88 as compared to in the case of original speech signals, which is consistent with the fact that children's speech spectra need greater compression to align with that of adults' speech. However, after pitch-normalization, though a larger number of signals have chosen optimal warp factors of 0.88 as compared to the original speech case, unlike in case of the speaking rate normalization, for all warp factors greater than 0.92 the number of normalized signals which have chosen those values have considerably increased compared to those in the original speech case. Further, we found few signals to choose warp factors >1 after pitch-normalization as against choosing values close to 0.88 in original speech case. Such estimation of warp factors after pitch-normalization is inappropriate as children's speech spectrum needs compression rather than expansion with respect to that of the adults' speech. This behavior is attributed to the distortions introduced in the spectra of few speech signals during explicit pitch-normalization as it involves decimation and/or interpolation for time-scaling operations. To overcome these distortions we have followed a "Back Off" procedure in which all those of warp factors estimated on the speaking rate and/or the pitch-normalized speech which have value greater than the ones obtained prior to that normalization are replaced by the values obtained on the signals prior to normalization. This allows us to reduce the errors introduced in the warp factor estimation particularly after pitch-normalization. The recognition performances on doing VTLN of the speaking rate and/or the pitch-normalized signals of the children test set "CH1" using the "Back Off" procedure are also given in the last three rows of Table 5 for the ease of comparison. It is noted that on using the warp factors estimated using the "Back Off" procedure for VTLN of the speaking rate and/or the pitch-normalized signals improves their recognition performances with significant improvements for signals involving the pitch-normalization. The combined normalization of the pitch, the speaking rate, and the formant frequencies of the signals results in a relative improvement of 80% over the baseline. This shows that the improvements obtained with the pitch and the speaking rate normalization are significant and additive to that obtained with VTLN.

5.2. Adults' Speech Recognition on Children's Speech Trained Models. Following the observations made in Section 5.1, it would be interesting to study the behavior of the normalization of the three identified significant acoustic correlates in context of adults' speech recognition on children's speech trained models. For this purpose, a new recognizer has been developed using the children training set "TR2" derived from the TIDIGITS corpus. The recognition performance for the children test set "CH2" and the adults test set "AD" on this new recognizer is 1.01% and 13.28%, respectively. It is to note here that the children's ASR performance is more than twice that of the adults' under matched condition. This is consistent with the known fact that children's speech has higher intraspeaker variability than adults' speech leading to larger variance of the acoustic models [13].

In our previous experiments we have noted that the normalization of the glottal parameters (OQ, RQ, SQ) do not have any significant effect on the ASR performance. Therefore, in this study we have explored the effect of normalization of the pitch, the speaking rate, and the formant frequencies only. For reducing the mismatch of

Table 6: Performances of the adult test set "AD" with and without normalization of different acoustic correlates. The 95% confidence interval for the performances is 0.64.

Condition	WER (%)
Baseline	13.28
Norm. (Formant Frequencies)	4.22
Norm. (Speaking Rate + Pitch + Formant Frequencies) (Using "Back Off" Procedure)	4.04

the adults' speech signals with respect to the children's speech trained acoustic models, various acoustic correlates have been modified appropriately in the same manner as described in Section 5.1. For ML-based normalization of various acoustic correlates of speech, the average pitch of the adults' test speech signals is transformed to seven different pitch values ranging from 160 Hz to 340 Hz with a step size of 30 Hz and the duration of the signals is increased by factors ranging from 1 to 1.39 with a step size of 0.075, thereby reducing the speaking rate of the signals by factors ranging from 1 to 0.7. For VTLN of the signals of the adults test set "AD", the ML search is performed among features warped by 13 equally spaced warping factors ranging from 0.9 to 1.14 with a step size of 0.02 for each signal. The baseline recognition performance of the adults test set "AD" along with those obtained with VTLN and the combined normalization of the pitch, the speaking rate, and the formant frequencies using the "Back Off" procedure (described in Section 5.1.5) is given in Table 6. From Table 6, it is noted that on combined normalization of the pitch, the speaking rate, and the formant frequencies of the adults' speech a relative improvement of 70% is obtained over the baseline for the adult test set "AD", which is comparable to the improvement obtained for the children test set "CH1" with children's speech normalization.

6. Conclusion

In this work, the effect of differences in various acoustic correlates of speech like the pitch, the speaking rate, the glottal parameters (OQ, RQ, SQ), and the formant frequencies for children's and adults' speech has been explored in the context of ASR under mismatched conditions. Our study done on a connected digit recognition task indicates that the differences in the pitch, the speaking rate, and the formant frequencies significantly affect the ASR performance and thus lead to significant improvement after normalization. On the other hand, the glottal parameters (OQ, RQ, SQ) have not been found to have any significant impact on the ASR performance. The normalization of the three significant acoustic correlates (the pitch, the speaking rate, the formant frequencies) in various combinations has also been studied. The experimental results show that we can successfully combine the improvements due to normalization of the above three acoustic correlates resulting in an overall relative improvement of 80% and 70% over the baseline for children's speech recognition and adults' speech recognition under mismatched conditions. Our future work aims at studying the effect of explicit normalization of the pitch and the speaking rate on the cepstral features like some studies are already relating the frequency warping for VTLN to the linear transformation of the cepstra. These cepstral domain transformations would not only ease the computational complexity of the normalization process but also would allow us to include the Jacobian factor of transformation in the estimation of the normalization factors.

Acknowledgment

Part of this work has been supported by the ongoing project No. SR/S3/EECE/39/2009 sponsored by the Department of Science and Technology, Government of India.

References

[1] P. Yildiz, "The multimedia interactive theatre by virtual means regarding computational intelligence in space design as HCI and samples from Turkey," *International Journal of Humanities and Social Sciences*, vol. 2, no. 1, 2008.

[2] D. Giuliani, O. Mich, and M. Nardon, "A study on the use of a voice interactive system for teaching English to Italian children," in *Proceedings of the 3rd IEEE International Conference on Advanced Learning Technologies (ICALT '03)*, pp. 376–377, July 2003.

[3] A. Hagen, B. Pellom, and R. Cole, "Children's speech recognition with application to interactive books and tutors," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU '03)*, pp. 186–191, December 2003.

[4] M. Russell, C. Brown, A. Skilling, et al., "Applications of automatic speech recognition to speech and language development in young children," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP '96)*, vol. 1, pp. 176–179, October 1996.

[5] M. Russell, R. W. Series, J. L. Wallace, C. Brown, and A. Skilling, "The STAR system: an interactive pronunciation tutor for young children," *Computer Speech and Language*, vol. 14, no. 2, pp. 161–175, 2000.

[6] D. Burnett and M. Fanty, "Rapid unsupervised adaptation to children's speech on a connected-digit task," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP '96)*, vol. 2, pp. 1145–1148, 1996.

[7] S. Narayanan and A. Potamianos, "Creating conversational interfaces for children," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 65–78, 2002.

[8] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: developmental changes of temporal and spectral parameters," *Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

[9] J. G. Wilpon and C. N. Jacobsen, "A study of speech recognition for children and the elderly," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '96)*, pp. 349–352, 1996.

[10] D. Giuliani and M. Gerosa, "Investigating recognition of children's speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, vol. 2, pp. 137–140, April 2003.

[11] G. Potamianos, S. Narayanan, and S. Lee, "Analysis of children speech: duration, pitch and formants," in *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech '97)*, pp. 473–476, 1997.

[12] M. Benzeguiba, R. D. Mori, O. Deroo, et al., "Automatic speech recognition and intrinsic speech variation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '06)*, vol. 5, May 2006.

[13] M. Gerosa, D. Giuliani, and F. Brugnara, "Acoustic variability and automatic recognition of children's speech," *Speech Communication*, vol. 49, no. 10-11, pp. 847–860, 2007.

[14] A. Potamianos and S. Narayanan, "Robust recognition of children's speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 603–616, 2003.

[15] M. Gerosa, D. Giuliani, and F. Brugnara, "Speaker adaptive acoustic modeling with mixture of adult and children's speech," in *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech '05)*, pp. 2193–2196, 2005.

[16] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–856, 1990.

[17] M. Iseli, Y.-L. Shue, and A. Alwan, "Age- and gender-dependent analysis of voice source characteristics," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '06)*, pp. I389–I392, 2006.

[18] B. Weinrich, B. Salz, and M. Hughes, "Aerodynamic measurements: normative data for children ages 6:0 to 10:11 Years," *Journal of Voice*, vol. 19, no. 3, pp. 326–339, 2005.

[19] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '96)*, vol. 1, pp. 353–356, May 1996.

[20] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP '96)*, pp. 1137–1140, 1996.

[21] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

[22] D. Giuliani, M. Gerosa, and F. Brugnara, "Improved automatic speech recognition through speaker normalization," *Computer Speech and Language*, vol. 20, no. 1, pp. 107–123, 2006.

[23] A. Potamianos, S. Narayanan, and S. Lee, "Automatic speech recognition for children," in *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech '97)*, pp. 2371–2374, Rhodes, Greece, September 1997.

[24] S. Das, D. Nix, and M. Picheny, "Improvements in children's speech recognition performance," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '98)*, pp. 433–436, 1998.

[25] A. Hagen, B. Pellom, and R. Cole, "Highly accurate children's speech recognition for interactive reading tutors using subword units," *Speech Communication*, vol. 49, no. 12, pp. 861–873, 2007.

[26] M. Benzeghiba, R. D. Mori, O. Deroo, et al., "Automatic speech recognition and speech variability: a review," *Speech Communication*, vol. 49, no. 10-11, pp. 763–786, 2007.

[27] J. Gustafson and K. Sjölander, "Voice transformations for improving children's speech recognition in a publicly available dialogue system," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP '02)*, pp. 297–300, September 2002.

[28] G. Stemmer, C. Hacker, S. Steidl, and E. Noth, "Acoustic normalization of children's speech," in *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech '03)*, pp. 1313–1316, 2003.

[29] J. P. Cabral and L. C. Oliveira, "Pitch-synchronous time-scaling for prosodic and voice quality transformations," in *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech '05)*, pp. 1137–1140, 2005.

[30] S. Young, G. Evermann, M. Gales, et al., *The HTK Book Version 3.4*, Cambridge University Engineering Department, Cambridge, UK, 2006.

[31] J. P. Cabral and L. C. Oliveira, "Pitch-synchronous time-scaling for high-frequency excitation regeneration," in *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech '05)*, pp. 1513–1516, 2005.

[32] R. Leonard, "A database for speaker-independent digit recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '84)*, pp. 42.11.1–42.11.4, 1984.

[33] R. Sinha and S. Ghai, "On the use of pitch normalization for improving children's speech recognition," in *Proceedings of the International Speech Communication Association (Interspeech '09)*, pp. 568–571, September 2009.

[34] J. Miller, "Effects of speaking rate on segmental distinctions," *Perspectives on the Study of Speech*, pp. 39–74, 1981.

[35] G. Peterson and I. Lehiste, "Duration of syllable nuclei in english," *Journal of the Acoustical Society of America*, vol. 32, pp. 693–703, 1960.

[36] T. Pfau, R. Faltlhauser, and G. Ruske, "A combination of speaker normalization and speech rate normalization for automatic speech recognition," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP '00)*, vol. 4, pp. 362–365, October 2000.

[37] N. Mirghafori, E. Fosler, and N. Morgan, "Towards robustness to fast speech in ASR," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '96)*, vol. 1, pp. 335–338, 1996.

[38] M. Siegler and R. Stern, "On the effects of speech rate in large vocabulary speech recognition systems," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '95)*, vol. 1, pp. 612–615, May 1995.

[39] "Open source software from the speech group," Wavesurfer version 1.8.5, January 2008, http://www.speech.kth.se/software.

[40] A. M. Sulter and H. P. Wit, "Glottal volume velocity waveform characteristics in subjects with and without vocal training, related to gender, sound intensity, fundamental frequency, and age," *Journal of the Acoustical Society of America*, vol. 100, no. 5, pp. 3360–3373, 1996.