A Study of Filter Bank Smoothing in MFCC Features for Recognition of Children's Speech

S. Umesh, *Member, IEEE*, and Rohit Sinha, *Member, IEEE*

***Abstract*—In this paper, we study the effect of filter bank smoothing on the recognition performance of children's speech. Filter bank smoothing of spectra is done during the computation of the Mel filter bank cepstral coefficients (MFCCs). We study the effect of smoothing both for the case when there is vocal-tract length normalization (VTLN) as well as for the case when there is *no* VTLN. The results from our experiments indicate that unlike conventional VTLN implementation, it is better *not* to scale the bandwidths of the filters during VTLN—only the filter center frequencies need be scaled. Our interpretation of the above result is that while the formant center frequencies may approximately scale between speakers, the formant bandwidths do not change significantly. Therefore, the scaling of filter bandwidths by a warp-factor during conventional VTLN results in differences in spectral smoothing leading to degradation in recognition performance. Similarly, results from our experiments indicate that for telephone-based speech when there is no normalization it is better to use uniform-bandwidth filters instead of the constant-Q like filters that are used in the computation of conventional MFCC. Our interpretation is that with constant-Q filters there is excessive spectral smoothing at higher frequencies which leads to degradation in performance for children's speech. However, the use of constant-Q filters *during* VTLN does not create any additional performance degradation. As we will show, during VTLN it is only important that the filter bandwidths are *not* scaled irrespective of whether we use constant-Q or uniform-bandwidth filters. With our proposed changes in the filter bank implementation we get comparable performance for adults and about 6% improvement for children both for the case of using VTLN as well as the for the case of not using VTLN on a telephone-based digit recognition task.**

***Index Terms*—Children's speech recognition, Mel filter bank, vocal-tract length normalization.**

I. Introduction

VOCAL-tract length normalization (VTLN) is used to improve the performance of speaker-independent speech recognition by accounting for interspeaker variability in the spectra for the same enunciated sound. One simple but effective technique to reduce the above variability is to warp the frequency axis of the utterance to compensate for the difference in formant positions between speakers [1]. This frequency-warping procedure was later efficiently implemented as a filter bank modification in Mel filter bank cepstral coefficients (MFCCs) feature front-end by Lee and Rose [2]. We refer to this approach as conventional VTLN. Recently, there have been many studies on the use of VTLN to improve the recognition performance of children's speech [3]–[6].

Automatic speech recognition (ASR) of children's speech has many applications particularly in the areas of education and entertainment. However, current ASR systems have been built mostly using only adult data because of easy availability of large adult speech databases. Such systems suffer significant degradation in recognition performance for children's speech. One of the main reasons for this performance degradation is the large spectral and temporal variability in children's speech when compared to adult speech for the same sound. Most of these variabilities are usually ascribed to the anatomical and morphological differences in the vocal tract. Hence, VTLN is used to reduce this acoustic variability and plays an important role in improving the recognition performance of children's speech.

One of the early studies on the recognition performance of children was done by Wilpon and Jacobsen [7]. They observed a sharp change in performance in the case of boys around puberty while the change was more gradual for girls. Burnett and Fanty [8] showed improvement for children's speech by appropriate shifts in psycho-acoustic scale to achieve speaker normalization. Potamianos *et al.* provided a detailed study of age-dependent effects on ASR performance and also the effect of different speaker-normalization schemes [4]. Some of the other studies on using VTL normalization for children's speech include [5], [6], and [9].

In this paper, we study the effect of scaling the filter bandwidths by frequency-warp-factor during VTLN (see Lee and Rose [2]). We also study the effect of spectral smoothing on the recognition performance of children's speech. In particular, we compare the performance using uniform-bandwidth filters with that of conventional filters that have increasing bandwidth with increasing frequency—which we refer to as constant-Q filters. In Section II, we review earlier studies on the age and gender-related variations in formant frequencies and bandwidths. Based on these studies, we point to the negative impact that the use of constant-Q like Mel filters and conventional VTLN bandwidth scaling of filters have on recognition performance of children's speech. In Section IV, we then discuss an alternate front-end signal processing that we have recently proposed [10], [11] and the improvement that it provides for children's speech when compared to conventional MFCC. Motivated by this, in [12], we have presented a series of experiments where we have changed different aspects of the front-end signal processing and reported

Manuscript received March 14, 2006; revised May 31, 2007. This work was supported in part by the Department of Science and Technology, Ministry of Science and Technology, India, under SERC project SR/S3/EECE/0008/2006. A part of this work was done while R. Sinha was with the Department of Electrical Engineering, Indian Institute of Technology, India. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Simon King.

S. Umesh is with the Department of Electrical Engineering, Indian Institute of Technology, Kanpur-208016, India (e-mail: sumesh@iitk.ac.in).

R. Sinha is with the Department of Electronics and Communication Engineering, Indian Institute of Technology, Guwahati-781039, India. (e-mail: rsinha@iitg.ernet.in).

Digital Object Identifier 10.1109/TASL.2007.906194

the corresponding effect on the recognition performance. In this paper, we analyze in detail the reasons for the differences in performance for the different front-ends. We have further expanded the set of experiments to include age- and gender-specific recognition results. Again, for each age or gender group we also analyze the differences in performance using the different front-ends. The analysis is done in Section VI and the age and gender-wise results are presented in Section VII. In Section VIII, we summarize the findings of this paper.

II. Variability in Formant Frequency and Bandwidth

There have been many studies that show that formant frequencies approximately scale among speakers enunciating the same sound [13]. Note that this is only an approximation, and there are significant deviations from this assumption [14]. However, in most VTLN studies, this approximate assumption is made, and we will also assume this model of linear scaling of formant frequencies for the purposes of this paper.

On the other hand, there have been very few studies on the relationship between formant bandwidths of speakers enunciating the same sound. Most formant bandwidth studies have looked at the variation of formant bandwidths between first, second, and higher order formants. Formant bandwidth studies by Dunn [15] and Fant [16], have shown that formants at higher frequencies have higher formant bandwidths. As Fant points out [17], formant bandwidths are typically of the order of 30–70 Hz for formants below 2000 Hz and increase at a higher rate above 2000 Hz. Investigations by Fujimura and Lindqvist [18] are one of the few studies that have done a detailed analysis of the formant bandwidths and the relationship of formant bandwidths *between* speakers. Some of their observations include the fact the female formant bandwidths are about 25% more than male formant bandwidths for *first* formant. However, the difference of the bandwidth values between male and female is seen to be *smaller* in the case of the second formant than for the first formant. Finally, for third formants there is no clear relation with the female formant bandwidths being even *lesser* than male formant bandwidths for some of the vowels. Although, their work does not discuss children formant bandwidth, we could extrapolate the male–female relationship to adult–children relationship. This would mean that while the first formant bandwidth of children may be higher than adults, the higher formant bandwidths are not significantly different from those of adults.

The above discussion indicates formant bandwidths do *not* scale among speakers for higher formants. This is in contrast to the assumption made in conventional VTLN that the entire spectra (including formant frequencies and bandwidths) are frequency-scaled versions of one and another for different speakers.

In conventional MFCC, the spectral smoothing is done using the triangular filters. These filters are uniformly spaced and have uniform bandwidth in the Mel domain. The use of Mel scale is motivated by perceptual studies and has been shown to perform better than linear frequency cepstrum, linear prediction cepstrum, linear prediction spectrum, or a set of reflection coefficients [19]. Therefore, in the linear-frequency domain the filters are nonuniformly spaced and have increasing bandwidth with increasing center frequency—and we refer to them as constant-Q-like filters in this paper. For example, for the 21 filters that we have used in the range of 200–3452 Hz for telephone speech, the first filter has a bandwidth of about 140 Hz and the 21st has a bandwidth of about 500 Hz. Note that for higher formants, although the formant frequencies of females (and children) are greater than those of males, the bandwidths are almost similar. Since the corresponding Mel filters at higher frequencies have wider bandwidth, there is more spectral-smoothing of higher formants of females (and children) when compared to those of males. We argue that this is responsible for the degradation in recognition performance of children when compared with adults. In this paper, we investigate the use of uniform-bandwidth filters instead of constant-Q filters and study the recognition performance for children's speech. Further, we also investigate whether it is beneficial to *not* scale the filter bandwidths during VTLN motivated by the study of Fujimura and Lindqvist that higher formant bandwidths do not scale among speakers. These modifications are also motivated by an alternate front-end called *WOSA-MFCC* that we recently introduced [10], [11], which shows improvement in recognition performance when compared to conventional MFCC for children's speech. We then analyze this difference in performance and argue that differences in spectral smoothing leads to loss in recognition performance in conventional VTLN. The recognition experiments are performed on a telephone-based connected digit recognition task whose setup is described next.

III. Experimental Setup

In this paper, since we are only considering the effect of the different front-ends on the recognition performance, a digit recognition task (with no complex language model) is sufficient to evaluate the efficacy of the features. The speech data for training the recognizer is derived from the Numbers corpus v1.0cd of OGI. The training set consists of 6078 utterances from adult male and female speakers. Two test sets are used: *adult* test set which is derived from Numbers corpus and consists of 2168 utterances from adult male and female speakers (other than training speakers) and *children* test set consisting of 2779 utterances from speakers ranging in age from 6 to 17 years. In the children's database (which is not publicly available), prior to data collection the (American) children were provided with instruction through their parents along with the speech material for the recording. A simple touch-tone interface was devised to automate the data collection. The data were manually verified for transcription accuracy. Both the adult and children test sets have utterances of variable digit string lengths. The adult test set consists of 789 male speakers and 1379 females speakers. The children test set consists of 1225 boys and 1554 girls. In the children test set, the age of 152 speakers were unavailable. There are 1547 children with ages above ten years and 1066 children with ages less than or equal to ten years. In the ten years and less category, there are 397 boys and 669 girls while in the above ten years category there are 736 boys and 811 girls. The word (digit) error rate is used to evaluate the performance of different methods.

The digit recognizer was developed using HTK HMM Toolkit. All the data used in this paper were obtained using a sampling frequency of 8000 Hz. We have used frame-size of

length 20 ms and a frame-rate of 100 Hz in all the experiments. The feature vector comprising of normalized energy, C_1 to C_{12} static cepstra and their first- and second-order derivatives is used, and cepstral mean subtraction is also performed. The digits are modeled as whole word simple left-to-right hidden Markov models (HMMs) without skips and have 16 states per word with five diagonal covariance Gaussian mixtures per state. The silence is modeled using three-state HMM model having six-mixture Gaussian models per state. A single-state short pause model tied to the middle state of the silence model is also used.

IV. Alternate Front-End for Normalization

We first present the alternate front-end referred to as *WOSA-MFCC feature*, where the spectral smoothing is done using a variant of averaged periodogram. As we will show, the WOSA-MFCC provides comparable performance for adults and superior performance for children when compared to conventional MFCC. The purpose of this study is to understand the cause for this difference in performance.

A. WOSA Smoothing Procedure

In ASR, the basic interest is in modeling the short-time spectral envelope of the speech. Therefore, almost all features used in ASR incorporate spectral smoothing in some form or another (one of the usual methods is by triangular filtering). For smoothing the short-time spectrum corresponding to speech frames, we have adopted a procedure known in literature as *weighted overlap segment averaging* (WOSA) [20] which is a variant of *averaged periodogram* spectral estimation method.

The steps involved in the WOSA procedure are as follows.

1) The given pre-emphasized speech frame (of length 20 ms) is segmented into a number of overlapping *subframes*.
2) Each subframe is Hamming windowed and the power spectrum is computed by squaring the magnitude of the Fourier transform of the windowed subframe.
3) A smoothed power spectrum is then obtained by averaging over the available subframes.

Since we usually use the Mel-warped spectrum, the autocorrelation estimates corresponding to that of averaged smoothed spectrum are first computed, and then the appropriate Mel-warped spectrum is derived by computing the nonuniform discrete Fourier transform (NDFT) of the averaged autocorrelation estimates. The subframe length and subframe overlap can be suitably chosen to control the amount of smoothing [21]. In this paper, the subframe length and overlap are chosen to be 80 and 60 samples, respectively (for sampling frequency of 8000 Hz), after extensive experimentation to get good performance for *both* adults and children.

B. Computation of WOSA-MFCC Feature and VTLN Warping

The NDFT in WOSA is computed at the *same* frequencies as the (nonuniformly spaced) center frequencies of Mel-scaled filter bank used in the case of unwarped the conventional MFCC feature computation. Finally, similar to MFCC feature computation, the above computed Mel-spaced spectrum is log compressed and converted to cepstral coefficients using DCT. Fig. 1 shows the block diagram of conventional MFCC and WOSA-MFCC front-ends, and it can be seen that they are *exactly* same in all respects except for the spectral smoothing procedure used.

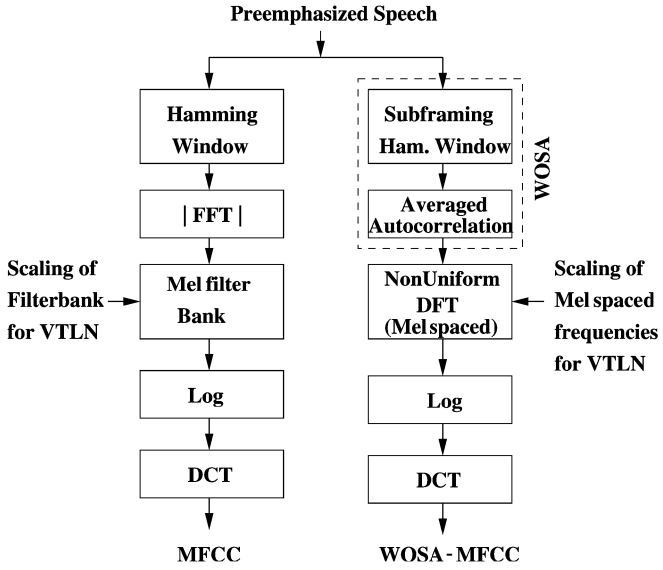

Fig. 1. Block diagram for the computation of conventional filter bank-based MFCC feature and our proposed WOSA-MFCC for a frame of speech.

TABLE I
Word Error Rate Using Conventional Filter Bank MFCC and Proposed WOSA-MFCC on a Connected Digit Recognition Task

Condition	Filter Bank Conventional MFCC		Averaged Periodogram WOSA-MFCC	
	Adults	Children	Adults	Children
No Normalization	3.16	15.20	3.26	14.26
VTLN	2.58	9.13	2.59	8.59

Note that in conventional MFCC, during VTLN warping either the power spectrum is frequency-warped [1] or equivalently the filter bank is scaled [2]. For implementing the VTLN warping in WOSA-MFCC, the NDFT is computed at frequencies that are obtained by appropriate VTLN scaling of the original Mel-spaced frequencies. Therefore, in this case too, the warped NDFT frequencies match the center frequencies of the warped filter bank of conventional MFCC.

C. Performance Comparison of Conventional and WOSA-MFCC

Table I shows the performance WOSA-MFCC feature when compared to the conventional MFCC features. Since the two front-ends are different, for each front-end we train the corresponding model with adult train data and test it with corresponding adult test data (different from train data) and children data. The first row of Table I shows the performance of the WOSA-MFCC and conventional MFCC for the "no-normalization" case. From the table it can be seen that while the adult performances are comparable, there is a 6.3% improvement in performance for children when using WOSA-MFCC. We then built the VTLN model as described in [2] and [11] and tested the VTLN performance for adult and children test speech. In this case, normalization is applied both during training and testing. In this paper, for normalization during training, a

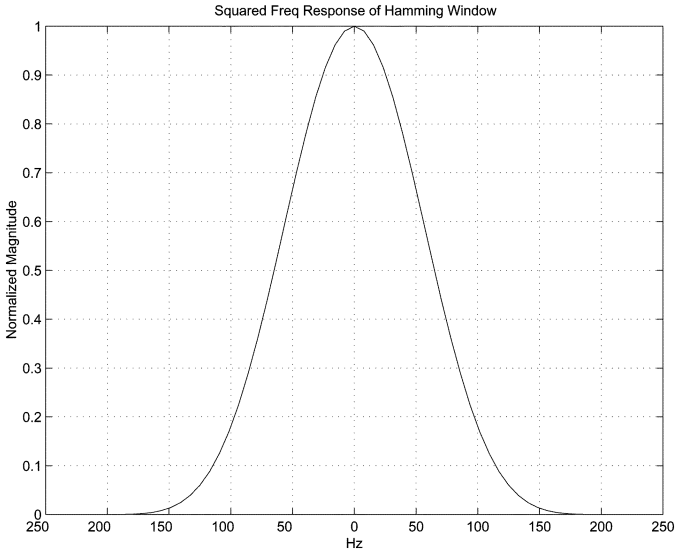

Fig. 2. Plot of normalized magnitude squared Fourier transform of Hamming window of length 80 samples.

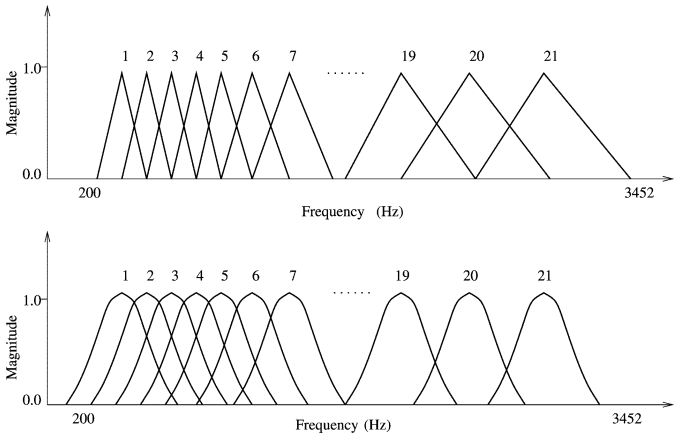

Fig. 3. Illustration of the explicit filter bank used in MFCC feature computation (top) and the implicit filter bank in WOSA-MFCC feature computation (bottom). Note that while center frequencies are identical, the bandwidths increase with increasing frequency in conventional MFCC while it is constant in WOSA-MFCC.

two-pass strategy is used since the correct transcription of the training data is available. On the other hand, during recognition, a full optimization approach is followed as described in [22]. On comparing the performance of VTLN using WOSA-MFCC feature with that of VTLN using conventional MFCC feature, we notice comparable performance for adults but a *significant* 6.4% improvement in performance for children. Therefore, when compared to conventional MFCC, WOSA-MFCC provides comparable performance for adults and about 6.3% improvement for children for both "no-normalization" and VTLN cases.

V. Filtering Interpretation of WOSA Procedure

In order to clearly understand the effect of the spectral smoothing procedure on the performance, the WOSA smoothing procedure (being a variant of averaged periodogram spectral estimation method) is interpreted as a filtering operation.

From Nuttall and Carter [20], in the WOSA method, the relationship between smoothed power spectra $\hat{G}_{av}(f)$ computed by Fourier transform of averaged autocorrelation estimate and the true power spectrum $G(f)$ of the speech frame can be expressed as follows:

$$\hat{G}_{av}(f) = G(f) * \{\mathcal{F}[w(t)]\}^2 \tag{1}$$

where "$*$" denotes convolution and $\mathcal{F}[w(t)]$ denotes Fourier transform of (Hamming) window used on each subframe.

So we can argue that, in the WOSA procedure, the smoothed power spectrum estimate at any frequency is obtained by bandpass filtering the true power spectrum at that frequency. The bandpass filter has a frequency response equal to that of square of Fourier transform of the Hamming window as shown in Fig. 2 and is called "WOSA-filter" in this work.

Since in the WOSA-MFCC feature, we have computed the power spectra at frequencies spaced on Mel scale, so it can be argued that the WOSA-based feature also uses a filter bank, similar to Mel-scaled filter bank used in MFCC feature computation, except that its constituent filters are of *uniform bandwidth* and have a frequency response equal to that of the WOSA filter, as shown at the bottom of Fig. 3. For ease of comparison, we have also shown in the figure (top part), the triangular Mel filter bank mask used in conventional MFCC feature computation. It can be seen that in conventional MFCC, the width of the filters or in turn the spectral smoothing is increasing with frequency unlike the the WOSA filter bank.

TABLE II
Recognition Performance of Conventional MFCC Using constant-Q Filter Bank and *WOSA-MFCC* and *Modified*-MFCC Using Uniform-Bandwidth Filter Bank

Conventional MFCC (Const.-Q filter bank) (Triang. Filter)		WOSA-MFCC (Uniform filter bank) (WOSA Filter)		Modified-MFCC (Uniform filter bank) (Triang. Filter)	
Adults	Children	Adults	Children	Adults	Children
3.16	15.20	3.26	14.26	3.15	14.41

VI. Study Into Effect of Spectral Smoothing

A. No VTL Normalization

We begin our study by considering the effect of filter bank smoothing when there is no normalization. The main differences in the filter banks of conventional MFCC and WOSA-MFCC are in their bandwidths, since their center frequencies are identical as seen in Fig. 3. The other difference is in the shape of the constituent filters—one is triangular while the other has the shape of the magnitude-squared Fourier transform of the Hamming window as shown in Fig. 2.

We first consider the effect of the use of constant-Q like bandwidth of the conventional MFCC when compared to the uniform bandwidth of WOSA-MFCC on the performance of the digit recognizer. As seen in Fig. 3, while the first few filters of the WOSA filter bank have wider bandwidth when compared

TABLE III
PERFORMANCE OF CONVENTIONAL MFCC USING CONSTANT-Q FILTER BANK AND *MODIFIED*-MFCC USING UNIFORM-BANDWIDTH FILTER BANK WHEN THERE IS *NO* VTL NORMALIZATION

Condition	Conventional MFCC (Const.-Q filter bank) (Triang. Filter)								Modified-MFCC (Uniform filter bank) (Triang. Filter)							
	Adults			Children					Adults			Children				
	All (2168)	Males (789)	Females (1379)	All (2779)	Boys (1225)	Girls (1554)	>10 yrs. (1549†)	≤10 yrs (1078†)	All (2168)	Males (789)	Females (1379)	All (2779)	Boys (1225)	Girls (1554)	>10 yrs. (1549†)	≤10 yrs. (1078†)
No-VTLN	3.16	3.96	2.73	15.20	10.38	19.07	8.20	24.48	3.15	4.23	2.56	14.41	9.63	18.18	7.83	23.18

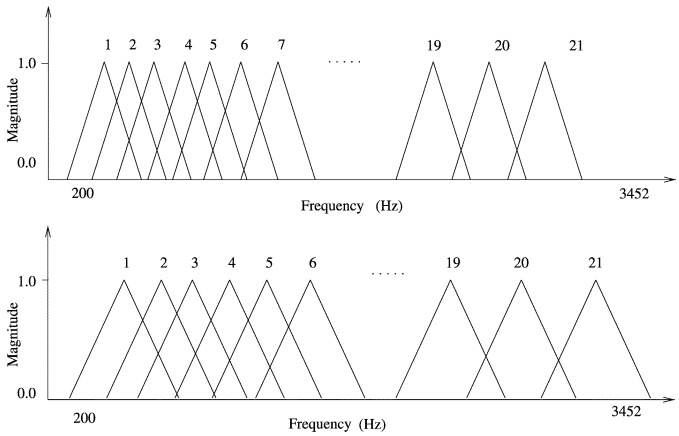

Fig. 4. Top figure shows the filter bank in *modified*-MFCC where the filters have uniform bandwidth unlike conventional MFCC. Bottom figure shows the bandwidth-scaled version of *modified*-MFCC during VTLN warping. Therefore, similar to conventional MFCC, both the bandwidth and the center frequency of the filters have been scaled by the VTLN warp factor.

to the filters in conventional MFCC, the higher frequency filters of WOSA have narrower bandwidth. This implies that there is increased smoothing and hence poorer resolution at higher frequencies for conventional MFCC. Since children's speech is concentrated mostly in the higher spectral region when compared to corresponding adult speech, children's speech is more affected by the loss of spectral resolution. As seen in Table II, the performance for children's speech improves by about 6% using the uniform-bandwidth filters when compared to the constant-Q like filters of conventional MFCC. Further, the performance for adults is comparable for both constant-Q and uniform-bandwidth filter banks.

Table II also shows that the performance does not change significantly irrespective of whether one uses triangular-shaped filter or the WOSA-filter of Fig. 2. This is seen by comparing the second and third column of the table. The filter bank used for producing the third column results is obtained by modifying the constant-Q filter with *uniform-bandwidth* triangular filters. We refer to this filter bank as *modified-MFCC*. The uniform-bandwidth filter bank-based *modified-MFCC* is obtained by modifying the constant-Q Mel filter bank used in conventional MFCC such that the constituent filters now have a *constant* bandwidth of 250 Hz (equal to the WOSA filter) while keeping their *triangular* shape and same center frequency. The three different filter banks used in Table II are shown in Fig. 3 and the top figure of Fig. 4. Therefore, the results indicate that using uniform BW filters (irrespective of shape) give comparable performance for adults and improved performance for children because of the improved resolution at higher frequencies when compared to the constant-Q filters of conventional MFCC.

Henceforth, we will compare the performance between conventional MFCC and modified-MFCC of Fig. 4 (and not WOSA-MFCC) since these two differ only in the bandwidths of the component filters. As we will show later, modified-MFCC and WOSA-MFCC have similar performance with and without VTLN indicating that the shape of the filters are not very important.

Table III shows the age-wise and gender-wise performance of modified-MFCC features and the conventional MFCC features for the case when there is *no* VTL normalization. First, we comment on the general trend in recognition performance across the different speaker categories, irrespective of the type of filter bank that we have used. Note that in the case of the no-normalization, the filter banks are left unchanged for the test speakers. Therefore, when we compare the outputs of the filter bank between an adult and a child speaker for the same spoken sound, there is a mismatch both due to the different formant frequencies and formant bandwidths. Recall that two-thirds of the training data were female adults. Therefore, the female test data are more closely matched to the model than the adult males or children and give the best performance. This is true for both the filter banks. Similarly, since boys have formant frequencies that are usually lower than those of girls [23], (and more closer to those of adults), their performance is better than that of girls. Similarly, the recognition performance of the above ten year olds (whose formants are much closer to adults) are much better than those of ten years and less.

We now compare the performance between the two filter banks. The difference between the two filter banks are in the bandwidth of the constituent filters. In modified-MFCC the filters have uniform bandwidth of 250 Hz while the conventional MFCC have bandwidths varying from 140 to 500 Hz. When compared to the MFCC filters, therefore, the uniform-bandwidth filters have wider bandwidths for frequencies below 1100 Hz and have narrower bandwidth for filters above 1100 Hz.

Since the lower formants have bandwidth in the range of 30–70 Hz for different speakers and higher formants have bandwidth around 250 Hz [17], there is more smoothing of lower formants when using uniform-bandwidth filters. Conversely, there is more smoothing of higher formants when using constant-Q filters. Therefore, each filter bank has over-smoothing in one half of the spectral range. For the female test data, there is comparable performance using the two different filter banks.

When there is no normalization of spectra, the dominant energies in spectra (i.e., formants) of males occur at lower frequencies than females for the same enunciated sound. From [18], we know that lower formants of males have narrower bandwidth than females (and children). During smoothing of spectra using triangle filters, the filters in modified-MFCC are much broader than conventional MFCC at low frequencies, resulting in excess smoothing of the male lower formants and consequent degradation in recognition performance as seen in the table.

Conversely, since the higher formants have similar bandwidth for all speakers, there is excessive smoothing of children spectra when using conventional MFCC filters, because the higher frequency filters of conventional MFCC are much broader than those of the modified-MFCC. This results in better performance for children when using modified-MFCC. This is true for both boys and girls and for both the age categories.

Since the age information was not available for some of the children, there are lesser number of speakers in the age-wise category, and therefore, we have used a † to emphasize this fact in Table III.

B. VTL Normalization

In this subsection, we study the effect of filter bank on the VTLN performance. The main idea in VTLN is to frequency-warp the spectra of a particular speaker so that the resulting spectra matches as closely as possible to the spectra of the *hypothetical* reference speaker used to train the VTL normalized acoustic model. This frequency-warp-factor or scale-factor is commonly denoted as α. It is normally assumed that α have values ranging from 0.80 to 1.20 based on physiological arguments. In conventional MFCC, a filter bank smoothing is applied on the spectra to obtain a smoothed spectra. Therefore, during VTLN instead of warping the spectra by α, a more efficient approach is to inverse-scale the filter bank while keeping the spectra unchanged as suggested in [2]. After normalization, the filter bank output of the test speaker should match as closely as possible to the filter bank output of the reference speaker. Note that in practice, instead of matching the spectra of test speaker to this "golden" or hypothetical reference speaker, we choose the warp-factor based on maximizing the likelihood with respect to the acoustic model. However, only for the purpose of easy visualization and understanding of the ideas presented in this paper, we will assume that the VTLN model is built using data from a hypothetical reference speaker. Further, although the filter bank output (and hence the spectral envelope) is used for feature computation, we will only talk about matching formant frequencies and formant bandwidths for ease of understanding. Throughout this paper, we will talk about formant frequencies and formant bandwidths when comparing the spectra between speakers. Similarly, we will talk about filter center frequencies and filter bandwidths when we talk about the filter bank used in VTLN.

In conventional VTLN during normalization, the entire filter bank is scaled by the warp factor. Therefore, not only are the filter center frequencies scaled but the corresponding filter bandwidths of the constituent filters are also scaled. The scaling of the filter center frequencies is equivalent to frequency warping which is necessary to align the formant frequencies to that of the hypothetical reference speaker. On the other hand, the scaling of the filter bandwidths implies the assumption that the corresponding formant bandwidth of the test speaker is also scaled by the same warp-factor. Therefore, conventional VTLN is based on the model assumption that *both* the formant frequencies and their bandwidths are scaled by the *same* warp-factor for any two speakers enunciating the same sound. An illustration of this assumption is shown in Fig. 5. The formants of the reference speaker are shown as "blobs" in Fig. 5(a). The formants of the test speaker are shown in dotted line in Fig. 5(b). Note that both the formant frequency and formant bandwidth of the test speaker are scaled version of the reference speaker. Correspondingly, for this assumption, the filter center frequencies and bandwidths are also scaled as seen in the Fig. 5(a) and (b). The output of each filter in the filter bank is obtained by multiplying the filter value with the corresponding spectral value at each discrete-frequency sample and summing them. In this case, the output of the ith filter of the hypothetical reference speaker [Fig. 5(a)] and that of suitably warped ith filter of test speaker [Fig. 5(b)] will be same except for possible constant gain-factor. Note that the use of "blobs" corresponding to the formants in spectra is to enable easier understanding of our arguments and does not correspond to any actual spectra.

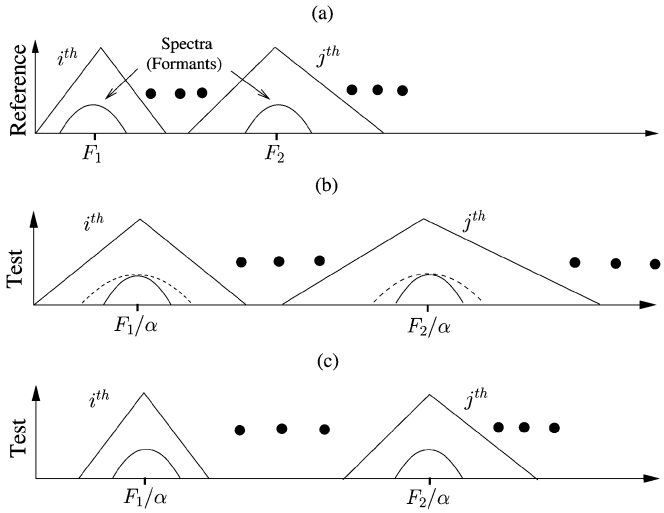

Fig. 5. Effect of filter bandwidth scaling in constant-Q filters used in conventional MFCC. The figure shows that if the formant bandwidth of the test speaker is scaled with respect to the reference speaker (shown in the dotted line), then the warped filter bank should have their bandwidth (and center frequency) scaled so that the two filter banks have similar output except for a constant gain. Conversely, as seen in (c) if formant bandwidth of test speaker is not scaled, the filter bandwidth should not be scaled. (a) Constant-Q Filter bank (b) Bandwidth and C.F. scaled constant-Q filter bank (c) Only C.F. scaled constant-Q filter bank.

In contrast to conventional VTLN, in the WOSA implementation of VTLN, *only* the filter bank center frequencies are scaled while the bandwidths are *not scaled*, i.e., the bandwidths are not altered and remain the same as in the unwarped case. Therefore, if there is *no* scaling of filter bandwidths during VTLN, then, this is equivalent to making the assumption that *only* the formant frequencies are scaled but not the formant bandwidths (which remain same as the unwarped case) for any two speakers enunciating the same sound. An illustration of this assumption is shown in Fig. 6(a) and (c) for the modified-MFCC case. Fig. 6(c)

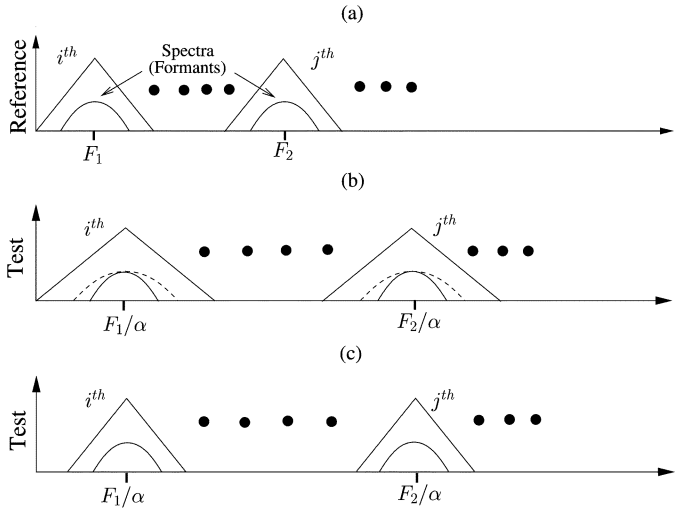

Fig. 6. Figures shows the relationship between the output of the filter banks of reference and test speakers for the case of uniform-bandwidth (but Mel-spaced) filters. As seen in (b), the filter bandwidth should be scaled if formant bandwidth is scaled (dotted line). On the other hand, if formant bandwidth is *not* scaled (solid line), then scaling of filter bandwidth leads to excessive smoothing. Finally, (c) shows the case of filter bandwidth and formant bandwidth not scaled. (a) Unfiorm-bandwidth filter bank. (b) Bandwidth and C.F. scaled uniform bandwidth with filter bank. (c) Only C.F. scaled uniform bandwidth filter bank.

corresponds to the case, where only the formant frequency of the test speaker is scaled, while the formant bandwidth is the *same* as the reference speaker. As seen from the figure, if we assume that the formant bandwidths are *not scaled*, then the filter bandwidths should also *not* be scaled (i.e., only the filter center frequencies should be scaled) in order for the warped filter bank outputs of test-speaker matches the filter bank outputs of the reference speaker. A similar analogy for the constant-Q case can be made by observing Fig. 5(a) and (c).

At this point, we would like to make some comments on the assumption of scaling of formant frequencies and the scaling/nonscaling of formant bandwidths. It is well known that the scaling of formant frequencies by a constant factor between two speakers enunciating the same sound is a crude assumption and there are significant deviations depending on the phone and the formant number [14]. Therefore, with this crude model, the formant frequencies will not be completely normalized. However, for the purposes of this paper, we will assume that this model is true. Further, although, neither the model of formant bandwidths being scaled nor the model that they are not at all scaled may be accurate, our hypothesis is that it is better to assume that the formant bandwidths are *not* scaled contrary to the assumption made in conventional VTLN. This is motivated by Fujimura and Lindqvist's analysis [18] which showed that although there is scaling of bandwidth for first formants, the bandwidth of higher formants are not significantly different among male and female speakers. A similar assumption about the bandwidth not scaling is also made in other areas of speech processing such as in estimation of source harmonics' magnitude [24] and in formant diphone parameter extraction [25]. We investigate the validity of this hypothesis by comparing the recognition performance on a digit recognition task by scaling/nonscaling the bandwidth during VTLN. In all the experiments based on VTLN, normalization is done during *training* and *testing*. The VTL normalized model is built from appropriately normalized adult data after iterative training. During testing, we test on VTL normalized adult and children data.

We now examine whether we need to scale the filter bandwidth during VTLN. Note that in all cases we will scale the filter center frequencies to match the formant frequencies between test and reference speakers. The two models that we will compare are as follows.

- In conventional VTLN, we make the model assumption that the formant bandwidths are scaled between two speakers enunciating the same sound. The scaled formant bandwidths are shown by dotted lines in Figs. 5(b) and 6(b). For this case, we would require the scaling of the bandwidths of the filters so that the output of the filter bank are same after normalization.
- Alternately, in our proposed model we assume that the formant bandwidths are *not* scaled between speakers enunciating the same sound. To be consistent with this assumption the filter bandwidths should not be scaled (only the filter center frequencies should be scaled), to ensure that the filter bank outputs are same after normalization as shown in Figs. 5(c) and 6(c).

These observations are true irrespective of whether we use constant-Q filters of conventional MFCC or the uniform-bandwidth filters of *modified-MFCC* in the filter bank. We now investigate which of these models are more appropriate especially for children's speech by comparing the normalization performance between scaling and not scaling the filter bandwidth.

1) Scaling of Bandwidth During VTLN: In this experiment during VTLN, we scale the filter center frequencies *and* the filter bandwidths by the *same* warp-factor. This is illustrated in Figs. 5(b) and 6(b). Note that during normalization we can think of the filter bank of the hypothetical reference speaker being appropriately moved to match the test spectra. Since the training data is female-dominated, the VTLN model is more closer to a hypothetical female reference speaker than a male reference speaker. Therefore, males have warp-factors greater than unity, females have warp-factors in the vicinity of unity and children have warp-factors smaller than unity.

The two filter banks under consideration have similar bandwidth of 250 Hz around 1100 Hz in the unnormalized case. Therefore, while lower frequency formants (those with frequencies less than 1100 Hz) have better resolution in constant-Q filter bank, the higher frequency formants have better resolution in uniform-bandwidth filters. As seen from Fig. 8 for constant-Q filters, the higher-frequency filters are very broad with bandwidth of more than 350 Hz and bandwidth scaling with $\alpha < 1$ makes them even broader resulting in excess smoothing and affecting children higher formants. Conversely, the narrow low-frequency constant-Q filters become narrower for $\alpha > 1$, helping improve the resolution of male lower formants. In the case of uniform-bandwidth filters, the low-frequency filters are broad with bandwidth being 250 Hz as seen in Fig. 7. Therefore, bandwidth scaling with $\alpha \ll 1$ makes them even broader resulting in excess smoothing affecting recognition performance of very small children.

TABLE IV
PERFORMANCE OF WARP-BASED SPEAKER NORMALIZATION METHOD USING CONVENTIONAL MFCC INVOLVING CONSTANT-Q FILTER BANK AND MODIFIED-MFCC INVOLVING UNIFORM-BANDWIDTH FILTER BANK

Condition	Conventional MFCC (Const.-Q filter bank) (Triang. Filter) {BW scaled}								Modified-MFCC (Uniform filter bank) (Triang. Filter) {BW scaled}							
	Adults			Children					Adults			Children				
	All (2168)	Males (789)	Females (1379)	All (2779)	Boys (1225)	Girls (1554)	>10 yrs. (1549*)	≤10 yrs (1078*)	All (2168)	Males (789)	Females (1379)	All (2779)	Boys (1225)	Girls (1554)	>10 yrs. (1549*)	≤10 yrs. (1078*)
No-VTLN	3.16	3.96	2.73	15.20	10.38	19.07	8.20	24.48	3.15	4.23	2.56	14.41	9.63	18.18	7.83	23.18
VTLN	2.58	3.00	2.35	9.13	7.02	10.80	5.41	13.50	2.48	3.09	2.14	9.18	6.86	11.01	5.28	14.04

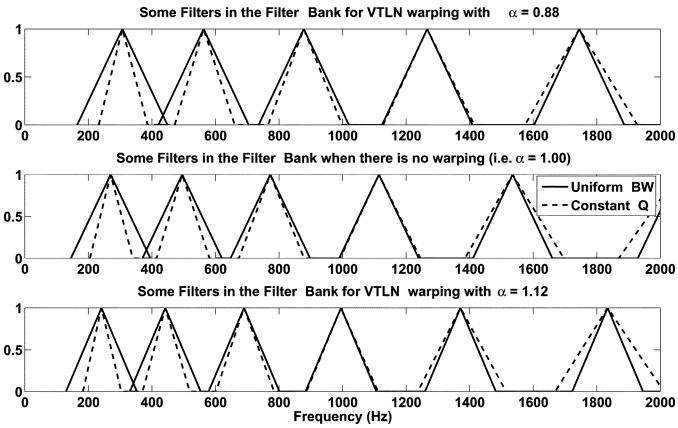

Fig. 7. Figure shows the *low-frequency* constant-Q filters and uniform-bandwidth filters for filter bandwidth scaling by different warp factors.

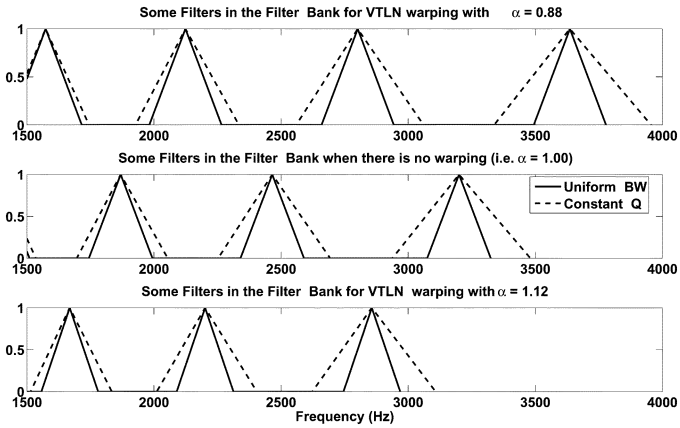

Fig. 8. Effect of filter bandwidth scaling for *high-frequency* constant-Q filters and uniform-bandwidth filters.

Based on the above observations, we analyze the performance shown in Table IV. Note that there is a always performance gain after normalization due to matching of the formant frequencies. In the following, we are analyzing the effect of filter bandwidths on the normalization performance.

1) In the *unnormalized* filter bank, the constant-Q filters have very wide filters at higher frequencies causing excessive smoothing of higher formants. Conversely, for low-frequencies, the uniform-bandwidth filters with bandwidth equal to 250 Hz (much greater than constant-Q bandwidth) cause excessive smoothing of lower formants. This is the reason that in the no-normalization case, males suffer degradation with uniform-bandwidth filters, while children have lower performance with constant-Q filters.
2) From Fujimura and Lindqvist studies [18], we know that male first formants have about 25% narrower bandwidth than corresponding female formant bandwidths. Hence, filter bandwidth scaling which results in narrower filters for $\alpha > 1$ helps in better matching the lower formants of the adult male test speakers for the case of constant-Q filters. This helps the male performance. On the other hand, for $\alpha < 1$ the already broad constant-Q filters at higher frequencies will become broader, resulting in excessive smoothing of the higher formants as seen in Fig. 8. This in turn affects the performance of children.
3) When uniform-bandwidth filters are used during normalization with bandwidth scaling, the uniform-bandwidth filters become narrower for $\alpha > 1$ similar to constant-Q case. However, they are still broader than corresponding constant-Q filters for lower formants. Therefore, for males, there is more smoothing using uniform-bandwidth filters when compared to constant-Q filters. As seen from Table IV, during normalization, males have inferior performance using uniform-bandwidth filters when compared to constant-Q filters.
4) We now consider the case of children test speakers. If we extrapolate Fujimura's findings to children, we may assume children formant bandwidths are the *same* as adult speakers (except possibly for first formants). With this assumption, during normalization, the scaling of the filter bandwidth with $\alpha < 1$ would result in more smoothing for children when compared to adult female speakers and more so with respect to adult male speakers. This can be seen for constant-Q filters in Fig. 5(b) and for uniform-bandwidth filters in Fig. 6(b). Since constant-Q filters are already broader than uniform bandwidth at higher frequencies, they will have more smoothing. This is partially responsible for degradation in normalization performance for boys and above-ten-year-olds when using constant-Q filters.
5) On the other hand, for the case of $\alpha \ll 1$, the already broad uniform-bandwidth filters (when compared to constant-Q) at low frequencies become even broader, resulting in excessive smoothing of *lower* formants of children. This has a negative influence on the performance when $\alpha \ll 1$ even though at higher frequencies, uniform-bandwidth filters

TABLE V
PERFORMANCE OF THE DIFFERENT FILTER BANKS WHEN FILTER BANDWIDTHS ARE NOT SCALED DURING VTLN WARPING

Condition	Conventional MFCC (Const.-Q filter bank) (Triang. Filter) {BW not scaled}								Modified-MFCC (Uniform filter bank) (Triang. Filter) {BW not scaled}							
	Adults			Children					Adults			Children				
	All (2168)	Males (789)	Females (1379)	All (2779)	Boys (1225)	Girls (1554)	>10 yrs. (1549*)	≤10 yrs (1078*)	All (2168)	Males (789)	Females (1379)	All (2779)	Boys (1225)	Girls (1554)	>10 yrs. (1549*)	≤10 yrs. (1078*)
No-VTLN	3.16	3.96	2.73	15.20	10.38	19.07	8.20	24.48	3.15	4.23	2.56	14.41	9.63	18.18	7.83	23.18
VTLN	2.58	3.09	2.30	8.53	6.86	9.84	5.07	12.47	2.57	3.30	2.17	8.69	6.90	10.09	5.42	12.97

provide better resolution than constant-Q filters. This explains the slight degradation in performance for girls and less-than-ten-year olds when using uniform-bandwidth filters.

The effect of excess smoothing due to bandwidth scaling seems to degrade the performance as seen in the case of girls and less-than-ten-year-olds for uniform-bandwidth filters. Similarly, the excess smoothing of the high-frequency constant-Q filters seem to degrade the performance of females, boys, and above-ten-year-olds. On the other hand, the narrowing of filter bandwidths due to scaling helps improve the resolution of lower formants improving the performance for males.

Although, the no-normalization performance is 15.22% for conventional MFCC and 14.43% for modified-MFCC, the children VTLN performance is comparable at about 9.15% for both the methods. This does *not* mean that there is less improvement using modified-MFCC. When there is no-normalization, conventional MFCC has poorer performance due to increased smoothing of the constant-Q filters at higher frequencies. On the other hand, during normalization, there is *no* extra smoothing for children in constant-Q filters, since after warping the corresponding filters of adults and children are the same as seen in Fig. 5. In other words, the filters of the hypothetical reference speaker are appropriately moved during normalization, and the outputs of *exactly* the same filters are compared for both adults and children. During normalization (assuming same warp-factor), the only difference between constant-Q and modified-MFCC filters are in their spectral resolution at different frequency regions. Therefore, both uniform-bandwidth and constant-Q filters have comparable normalization performance for children as seen from Table IV.

In the above experiment, we have scaled the filter bandwidth during VTL normalization. In most cases, there is a degradation in performance whenever there is excessive spectral smoothing due to filter bandwidth scaling. It would, therefore, be useful to improve upon the normalization performance by avoiding the excessive smoothing that results from bandwidth scaling. We have argued using Figs. 5(c) and 6(c) that if the formant bandwidths are *not* scaled for different speakers enunciating the same sound, then the filter bandwidths should not be scaled during VTLN. Since this is approximately true except for first formants, we should expect an improvement in performance when we repeat the experiment by not scaling the bandwidth. These sets of experiments are discussed in Section VI-B2.

2) Bandwidth Not Scaled During Warping: We now study the effect of scaling only the filter center frequency but *not* the filter bandwidth during VTLN warping. Therefore, this approach should avoid the excessive smoothing of constant-Q filters due to bandwidth scaling for children. Similarly, for $\alpha \ll 1$, the excessive smoothing of uniform-bandwidth filters for lower formants of children will also be avoided. On the other hand, the narrowing of bandwidth for $\alpha > 1$ will not take place in this case, and hence adult males have some degradation in performance. We discuss below the effect of bandwidth not being scaled on the normalization performance for various genders and age groups by comparing Tables IV and V.

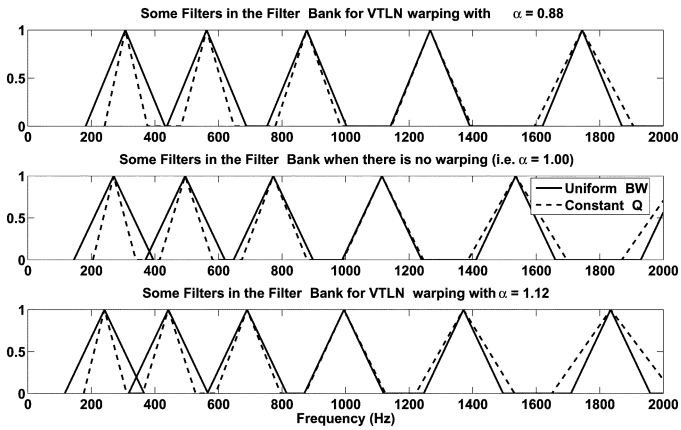

Fig. 9. Figure shows the *low-frequency* constant-Q filters and uniform-bandwidth filters when there is no filter bandwidth but only center frequency scaling.

1) Since there is no bandwidth scaling, the filters of the hypothetical reference speaker (more like female for this database) are just appropriately moved to match the test speaker. Therefore, for male test data the filters are slightly broader than for the bandwidth-scaling case. Since the first formants of males are narrower than females but the bandwidth of filters remain same, there is some extra smoothing of lower formants when compared to the bandwidth-scaling case. Therefore, we would expect some degradation in performance of males when compared to the bandwidth-scaling case. This is seen to be true for both constant-Q filters and uniform-bandwidth filters by comparing Tables IV and V. The degradation is more for uniform-bandwidth filters than constant-Q since they are they are much broader at low frequencies resulting in excess smoothing (see Fig. 9).
2) In the case of uniform-bandwidth filters, when there is no bandwidth-scaling, the excessive smoothing of the lower

formants of children (with $\alpha \ll 1$) is avoided. Further, since the higher formant bandwidths are similar among children and adults, all speakers have similar smoothing when there is no-bandwidth-scaling avoiding bandwidth mismatch that is present in the bandwidth-scaling case. Therefore, girls and less-than-10-year olds benefit from no-bandwidth-scaling when compared to the bandwidth-scaling case.

3) In the case of constant-Q filters, for $\alpha \ll 1$ the narrower (when compared to bandwidth-scaling case) lower-frequency filters are appropriately moved to match the children spectra. If we assume that the first formant of children are broader, then there is less smoothing and therefore some bandwidth mismatch. On the other hand, since the higher-frequency filters are not scaled, the excessive smoothing of higher-formants is avoided resulting in better performance when compared to bandwidth-scaling case (see Fig. 10). Therefore, again, girls and less-than-10-year olds benefit from no-bandwidth-scaling.
4) Note that both boys and above-ten-year-olds have some speakers whose α are close to that of adult males, i.e., $\alpha > 1$. Note that boys above ten years old are common to both groups. In the case of constant-Q filters for speakers with $\alpha > 1$, when there is no-bandwidth-scaling, the lower frequency filters are broader when compared to bandwidth-scaling case, and hence there is some loss in resolution. This may marginally degrade performance. On the other hand for speakers with $\alpha < 1$, the broad higher-frequency filters are not scaled and hence they have better resolution. This helps speakers in the group who have $\alpha < 1$ and gives some improvement in performance. The net effect for boys and above-ten-year-olds is that there is some improvement in performance for the case of no-bandwidth-scaling.
5) In the case of uniform-bandwidth filters, the bandwidth scaling of the higher frequency filters may not significantly affect performance when compared to not scaling. However, the bandwidth scaling of lower frequency, broad, uniform-bandwidth filters might affect the performance for the case of $\alpha < 1$. On the other hand, the narrowing of the low-frequency filters for $\alpha > 1$ helps provide better resolution. The net effect seems to be comparable performance between bandwidth-scaling and no-bandwidth-scaling for uniform-bandwidth filters for the case of boys and above-ten-year-olds.
6) Similarly, for adult females (with α mostly around unity), the performance seems to be comparable between bandwidth-scaling and not-scaling cases for both constant-Q and uniform-bandwidth filters.

We now compare the performance between constant-Q filters and uniform-bandwidth filters for the case of no-bandwidth-scaling as shown in Table V. It is important to remember that during normalization, the filter bank corresponding to the hypothetical reference speaker (who is close to a female) is appropriately moved with no change in bandwidth in this case. The difference in performance may be analyzed as follows.

1) The better resolution of high-frequency filters for the uniform-bandwidth case seems to help the female speakers (with $\alpha \approx 1$) marginally when compared to the constant-Q case.
2) Since the bandwidths of higher formants are similar among different speakers, the corresponding normalized filter bank outputs are similar at high frequencies. This is seen from Figs. 5(c) and 6(c). Therefore, there is no bandwidth mismatch with respect to the hypothetical reference speaker both for the case of uniform-bandwidth and constant-Q filter case.
3) On the other hand for lower formants, in the case of constant-Q filters, the narrower (when compared to the unnormalized) filters of the hypothetical reference speaker are moved right for children with $\alpha < 1$ resulting in better resolution. The corresponding uniform-bandwidth filters have constant bandwidth of 250 Hz which is much broader than the constant-Q filter, and hence there is loss in resolution. This results in degradation in performance for uniform-bandwidth filters as seen for girls and less-than-ten-year olds.
4) In the case of $\alpha > 1$, the filters are moved to the left, and hence the constant-Q filters become broader with respect to the original *unnormalized* filter at that frequency. However, the corresponding uniform-bandwidth filters are still much broader resulting in loss of resolution in this case too. Hence, there is a degradation in performance for males too when using uniform-bandwidth filters.
5) Similarly, boys and above-ten-year-olds have marginal gain using constant-Q filters without bandwidth-scaling when compared to the uniform-bandwidth filters.

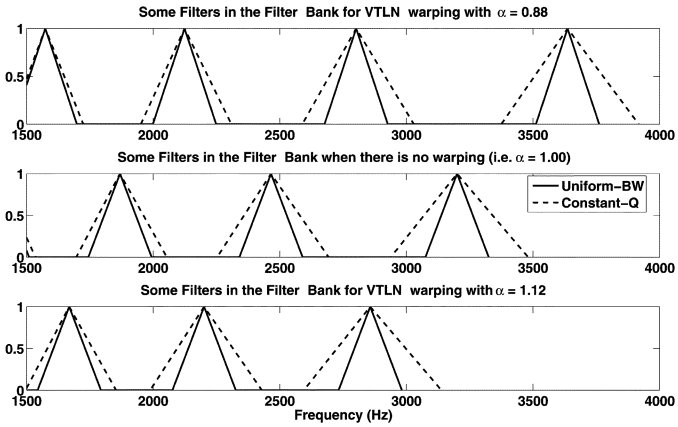

Fig. 10. Effect of No filter bandwidth-scaling for high-frequency constant-Q filters and uniform-bandwidth filters.

Note that *both* constant-Q and uniform-bandwidth filters have better performance for children when there is no-bandwidth-scaling as compared to bandwidth-scaling case. When we compare the performance for children in Tables IV and V, we see that there is about 7% relative improvement obtained by not scaling the filter bandwidth during VTLN. In the case of no-bandwidth-scaling, constant-Q filters are marginally better than uniform-bandwidth filters for children as seen in Table V.

From the above experiments, we conclude that the bandwidth of the filter should not be scaled during normalization of children's speech.

TABLE VI
PERFORMANCE OF VTLN USING CONVENTIONAL MFCC FILTER BANK INVOLVING CONSTANT-Q FILTERS. THE FILTER BANDWIDTHS ARE SCALED DURING NORMALIZATION

	Conventional MFCC (Const.-Q filter bank) (Triang. Filter) {BW scaled}											
Condition	All Children				Boys				Girls			
	6–9	10–11	12–13	≥ 14	6–9	10–11	12–13	≥ 14	6–9	10–11	12–13	≥ 14
No-normalization	27.48	15.59	6.43	6.39	19.15	12.35	6.29	4.14	32.35	17.74	6.59	8.34
VTLN	15.57	8.79	3.77	4.80	11.04	6.99	4.33	3.82	18.21	9.99	3.11	5.66

TABLE VII
PERFORMANCE OF VTLN OF MODIFIED-MFCC USING UNIFORM-BANDWIDTH FILTERS. THE FILTER BANDWIDTHS ARE SCALED DURING NORMALIZATION

	Modified-MFCC (Uniform filter bank) (Triang. Filter) {BW scaled}											
Condition	All Children				Boys				Girls			
	6–9	10–11	12–13	≥ 14	6–9	10–11	12–13	≥ 14	6–9	10–11	12–13	≥ 14
No-normalization	26.18	14.47	6.20	6.25	17.93	10.61	5.88	4.14	30.99	16.65	6.59	8.07
VTLN	16.21	8.47	3.88	4.62	11.75	6.88	4.12	4.06	18.81	9.53	3.59	5.10

TABLE VIII
PERFORMANCE OF VTLN OF MODIFIED-MFCC USING UNIFORM-BANDWIDTH FILTERS. THE FILTER BANDWIDTHS ARE NOT SCALED DURING NORMALIZATION

	Modified-MFCC (Uniform filter bank) (Triang. Filter) {BW not scaled}											
Condition	All Children				Boys				Girls			
	6–9	10–11	12–13	≥ 14	6–9	10–11	12–13	≥ 14	6–9	10–11	12–13	≥ 14
No-normalization	26.18	14.47	6.20	6.25	17.93	10.61	5.88	4.14	30.99	16.65	6.59	8.07
VTLN	14.97	8.14	3.99	4.92	10.74	8.16	4.74	3.66	17.45	8.06	3.11	5.86

VII. ANALYSIS OF CHILDREN RECOGNITION PERFORMANCE

In this section, we analyze the performance of children for different age categories and separately for boys and girls. Since the vocal-tract length increases as a child grows until it reaches adult size, the performance of children improves with increasing age when we use models that have been trained with adult data. The only exception are females who display variable patterns of increases and decreases in the warp-factor values for different vowels after the age of 15 years as discussed by Whiteside [26]. As studied by Whiteside in detail, females display a decrease in variation of first formant frequencies between prepuberty to puberty, with subsequent increases between postpuberty and adulthood. Further, formant frequencies of the 18-year-old females are significantly higher than those values for the adult females [26]. In the case of females, therefore, there is degradation in performance for the above-14-year olds when compared to the 12–13 age group.

Note that, since boys have lower formant frequencies than girls (i.e., more closer to adults) for the same age, we see that in any age category, the performance of boys is better than those of girls. Tables VI–IX show the performance of constant-Q and uniform-bandwidth filter banks for the cases of bandwidth-scaling and no-bandwidth-scaling respectively. The following observations can be made from the tables.

- When there is no VTL normalization, the performance is better when we use uniform-bandwidth filters when compared to constant-Q filters. This is true for all age groups and for both boys and girls.
- In the case of uniform-bandwidth filters with bandwidth-scaling during VTLN, the performance of boys in the 10–11 and 12–13 category are much better than the corresponding case when there is no-bandwidth-scaling. This may be due to the fact that the slight increase in formant bandwidth of lower formants (when compared to the adult hypothetical speaker) is not accounted for in the no-bandwidth-scaling case. This results in slight differences in filter bank outputs.
- On the other hand, for boys in the age group 6–9 years, who are mostly children with $\alpha < 1$, there is improvement when there is no-bandwidth-scaling in the case uniform-bandwidth filters. This may be due to the fact that the excessive smoothing of lower formants is avoided.
- A similar behavior also observed in the case of constant-Q filters, where the performance of 10–11 and 12–13 year-olds are better when there is bandwidth

TABLE IX
PERFORMANCE OF VTLN USING CONVENTIONAL MFCC FILTER BANK INVOLVING CONSTANT-Q FILTERS. THE FILTER BANDWIDTHS ARE NOT SCALED DURING NORMALIZATION

	Conventional MFCC (Const.-Q filter bank) (Triang. Filter) {BW not scaled}											
Condition	All Children				Boys				Girls			
	6–9	10–11	12–13	≥ 14	6–9	10–11	12–13	≥ 14	6–9	10–11	12–13	≥ 14
No-normalization	27.56	15.59	6.48	6.39	19.25	12.35	6.29	4.14	32.41	17.74	6.71	8.34
VTLN	14.56	7.72	3.77	4.43	10.94	7.69	4.74	3.42	16.68	7.75	2.63	5.31

scaling. Since the constant-Q filters are much narrower than uniform-bandwidth filters and further since boys are physiologically more closer to adult females, there is not much difference in performance for 6–9 year olds between the cases of bandwidth-scaling and no-bandwidth-scaling.

- In both the constant-Q and uniform-bandwidth filters, bandwidth *not* scaling seems to provide improved performance when compared to the bandwidth-scaling case for the above-14-year-old boys.
- In the case of uniform-bandwidth filters, we get better performance when bandwidth is *not* scaled during VTLN for all age-categories of girls except those above 14 years.
- Finally, for the case of constant-Q filters, bandwidth *not* scaling provides improvement for all age categories in girls.

Therefore, except for boys in the 10-11 and 12–13 year-old categories, in all other age-categories of *both* boys and girls, the use of constant-Q filters with *no* bandwidth scaling seems to provide the best VTL normalization performance among the different front-ends that we have considered.

VIII. SUMMARY

Our experiments on a telephone-based connected-digit recognition task indicate that it is important that filter bandwidths are *not* scaled during VTLN to get improved normalization performance especially for children. These observations can be properly interpreted if we assume that the formant bandwidths remain approximately the same and only the formant frequencies are scaled when two speakers enunciate the same sound. Formant-bandwidth studies by Fujimura and Lindqvist [18] show that the above assumption is approximately true except for the first formant. Our experiments also indicate that at least for telephone speech there is no significant difference in normalization performance between constant-Q and uniform-bandwidth filters for adults. This is because while constant-Q filters have better resolution at lower frequencies, uniform-bandwidth filters have better resolution at higher frequencies. Therefore, adult males benefit from the use of constant-Q filters, while adult females benefit from the use of uniform-bandwidth filters.

On the other hand, when there is no normalization, since children's speech have mostly higher frequency components when compared to adults, there is excessive smoothing when we use constant-Q filters. This results in degradation in performance when compared with uniform-bandwidth filters. This problem does not arise during VTLN, since after warping the corresponding filters for adults and children are the same and there is no extra smoothing due to constant-Q property for children as seen in Fig. 5(c).

Therefore, using adult models if we wish to get improved recognition performance for children with and without using VTLN, we suggest the following implementation of the filter bank.

- When recognition of children's speech is done without any normalization (i.e with unnormalized adult models), it is better to use uniform-bandwidth filters (during training and testing) when compared to the constant-Q filters. This is based on the argument that the spectra for children are mostly concentrated in the higher frequency regions when compared to the adult spectra. Since the constant-Q frequency filters have broader filters at higher-frequencies, there will be loss of spectral resolution when compared to uniform-bandwidth filters.
- When VTLN is used during training and testing, then it is preferable to *only* scale the filter center frequencies and not the filter bandwidths. This approach, when compared to the conventional approach, gives better VTLN performance both for constant-Q and uniform-bandwidth filters in most cases for children. Further, when compared to uniform-bandwidth filters, the use of constant-Q filters gives marginally better performance.

We conclude that during VTLN, the bandwidth of the MFCC filters should *not* be scaled, only the center frequencies should be scaled to get improved performance for children.

ACKNOWLEDGMENT

The authors would like to especially thank D. R. Sanand for helping them run many of the experiments, and they would also like to thank S. Rath for his help in running the experiments in Section VII. The authors would also like to thank the anonymous reviewers for their many useful comments and suggestions which have helped improve the manuscript.

REFERENCES

[1] A. Andreou, T. Kamm, and J. Cohen, "Experiments in vocal tract normalization," in *Proc. CAIP Workshop: Frontiers in Speech Recognition II*, 1994.

[2] L. Lee and R. Rose, "Frequency warping approach to speaker normalization," *IEEE Trans. Speech, Audio Process.*, vol. 6, no. 1, pp. 49–59, Jan. 1998.

[3] A. Potamianos and S. Narayanan, "Creating conversational interfaces for children," *IEEE Trans. Speech, Audio Process.*, vol. 10, no. 1, pp. 65–78, Jan. 2002.

[4] A. Potamianos and S. Narayanan, "Robust recognition of children's speech," *IEEE Trans. Speech, Audio Process.*, vol. 11, no. 6, pp. 603–616, Nov. 2003.

[5] D. Giuliani and M. Geroso, "Investigating recognition of children's speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Hong Kong, Apr. 2003, pp. 137–140.
[6] M. Blomberg and D. Elenius, "Comparing speech recognition for adults and children," in *Proc. Fonetik*, Stockholm, Sweden, 2004, pp. 156–159.
[7] J. G. Wilpon and C. N. Jacobsen, "A study of speech recognition for children and the elderly," in *Proc. IEEE Int. Conf. Acoustic, Speech, Signal Process.*, May 1996, vol. 1, pp. 349–352.
[8] D. C. Burnett and M. Fanty, "Rapid unsupervised adaptation to children's speech on a connected-digit task," in *Proc. Int. Conf. Spoken Lang. Process.*, Philadelphia, PA, 1996, pp. 1145–1148.
[9] S. Das, D. Nix, and M. Picheny, "Improvements in children's speech recognition performance," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Seattle, WA, May 1998, pp. 433–436.
[10] S. Umesh, L. Cohen, N. Marinovic, and D. Nelson, "Scale transform in speech analysis," *IEEE Trans. Speech, Audio Process.*, vol. 1, no. 1, pp. 40–45, Jan. 1999.
[11] R. Sinha and S. Umesh, "Non-uniform scaling based speaker normalization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Orlando, FL, May 2002, vol. 1, pp. 589–592.
[12] S. Umesh, R. Sinha, and S. V. B. Kumar, "An investigation into front-end signal processing for speaker normalization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Montreal, QC, Canada, 2004, pp. 345–348.
[13] S. Lee, A. Potamianos, and S. Narayanan, "Acoustic of children's speech: Developmental changes of temporal and spectral parameters," *J. Acoust. Soc. Amer.*, vol. 105, no. 3, pp. 1455–1468, 1999.
[14] G. Fant, "A non-uniform vowel normalization," *STL-QPSR*, no. 2–3, pp. 1–19, 1975.
[15] H. K. Dunn, "Methods of measuring vowel formant bandwidths," *J. Acoust. Soc. Amer.*, vol. 33, no. 12, pp. 1737–1746, 1961.
[16] G. Fant, "Formant bandwidth data," *STL-QPSR*, no. 1, pp. 1–2, 1962.
[17] G. Fant, "Vocal tract wall effects, losses, and resonance bandwidths," *STL-QPSR*, no. 2–3, pp. 28–58, 1972.
[18] O. Fujimura and J. Lindqvist, "Sweep-tone measurements of vocal-tract characteristics," *J. Acoust. Soc. Amer.*, vol. 49, pp. 541–558, 1971.
[19] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuous spoken sentences," *IEEE Trans. Acoust., Speech Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.
[20] A. H. Nuttall and G. C. Carter, "Spectral estimation using combined time and lag weighting," *Proc. IEEE*, vol. 70, no. 9, pp. 1115–1125, Sep. 1982.
[21] R. Sinha, "Front-end signal processing for speaker-normalization," Ph.D. dissertation, Indian Inst. Technol., Kanpur, India, 2004.
[22] L. Welling, H. Ney, and S. Kanthak, "Speaker adaptive modeling by vocal tract normalization," *IEEE Trans. Speech, Audio Process.*, vol. 10, no. 6, pp. 415–426, Sep. 2002.
[23] P. A. Busby and G. L. Plant, "Formant frequency values of vowels produced by preadolescent boys and girls," *J. Acoust. Soc. Amer.*, vol. 97, no. 4, pp. 2603–2606, 1995.
[24] M. Iseli and A. Alwan, "An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Montreal, QC, Canada, May 2004, pp. 669–672.
[25] R. H. Mannell, "Formant diphone parameter extraction utilizing a labeled single-speaker database," in *Proc. Int. Conf. Spoken Lang. Process.*, Sydney, Australia, 1998, pp. 2003–2006.
[26] S. P. Whiteside, "Sex-specific fundamental and formant frequency patterns in a cross-sectional study," *J. Acoust. Soc. Amer.*, vol. 110, no. 1, pp. 464–478, 2001.

S. Umesh (M'93) received the Ph.D. degree in electrical engineering from the University of Rhode Island, Kingston, in 1993.

From 1993 to 1996, he was a Postdoctoral Fellow at the City University of New York. Since 1996 he has been with the Indian Institute of Technology, Kanpur, where he is currently an Associate Professor of Electrical Engineering. He has also been a Visiting Researcher at AT&T Research Laboratories, Florham Park, NJ, the Machine Intelligence Laboratory, Cambridge University, Cambridge, U.K., and the Department of Computer Science (Lehrstuhl für Informatik VI), RWTH-Aachen, Aachen, Germany. His recent research interests have been mainly in the area of speaker normalization and acoustic modeling and their application in large-vocabulary continuous-speech recognition systems. He has also worked in the areas of statistical signal processing and time-varying spectral analysis.

Dr. Umesh is a recipient of the Indian AICTE Career Award for Young Teachers in 1997 and the Alexander von Humboldt Research Fellowship in 2004.

Rohit Sinha (M'06) received the M.Tech. and Ph.D. degrees in electrical engineering from the Indian Institute of Technology, Kanpur, in 1999 and 2005, respectively.

From 2004 to 2006, he was a Postdoctoral Researcher at Cambridge University, Cambridge, U.K., in the Machine Intelligence Laboratory. Since 2006, he has been with the Indian Institute of Technology, Guwahati, as an Assistant Professor in the Department of Electronics and Communication Engineering. His research interests include speech and speaker recognition, noise robust speech processing, and audio segmentation.