Circuits, Systems, and Signal Processing (2019) 38:4667–4682
https://doi.org/10.1007/s00034-019-01072-7

Role of Linear, Mel and Inverse-Mel Filterbanks in Automatic Recognition of Speech from High-Pitched Speakers

Hemant Kumar Kathania[1] · S. Shahnawazuddin[2] · Waquar Ahmad[3] · Nagaraj Adiga[4]

Received: 9 July 2018 / Revised: 17 February 2019 / Accepted: 18 February 2019 /
Published online: 26 February 2019

Abstract

In the context of automatic speech recognition (ASR), the power spectrum is generally warped to the Mel-scale during front-end speech parameterization. This is motivated by the fact that human perception of sound is nonlinear. The Mel-filterbank provides better resolution for low-frequency contents, while a greater degree of averaging happens in the high-frequency range. The work presented in this paper aims at studying the role of linear, Mel and inverse-Mel-filterbanks in the context of ASR. When speech data are from high-pitched speakers like children, there is a significant amount of relevant information in the high-frequency region. Hence, down-sampling the information in that range through Mel-filterbank reduces the recognition performance. On the other hand, employing inverse-Mel or linear-filterbanks is expected to be more effective in such cases. The same has been experimentally validated in this work. For that purpose, an ASR system is developed on adults' speech and tested using data from adult as well as child speakers. Significantly improved recognition rates are noted for children's as well adult females' speech when linear or inverse-Mel-filterbank is used. The use of linear filters results in a relative improvement of 21% over the baseline. To further boost the performance, vocal-tract length normalization, explicit pitch scaling and pitch-adaptive spectral estimation are also explored on top of linear filterbank.

Keywords Automatic speech recognition · Children's speech recognition · Linear filterbank · VTLN · Pitch scaling

✉ S. Shahnawazuddin
s.syed@nitp.ac.in

Extended author information available on the last page of the article

1 Introduction

The task of developing an automatic speech recognition (ASR) system can be broken down into three major components, namely front-end speech parameterization, training acoustic models and language modeling. In this paper, the primary focus is on the first component, i.e., front-end speech parameterization. The basic motivation for front-end speech parameterization is to derive a compact representation for raw speech waveform after discarding the irrelevant information. In the context of ASR, the speaker- and environment-dependent acoustic attributes need to be suppressed. Consequently, the ASR system becomes speaker and ambiance independent. In addition to that, since the raw speech data are represented in a compact manner, the overall complexities of training system parameters as well as network search and decoding are reduced significantly.

A number of techniques have been proposed over the years for extracting front-end features from raw speech data. Among those, the one based on Mel-frequency cepstral coefficients (MFCC) [7] has been the dominant one. During MFCC feature extraction process, the short-time magnitude/power spectra of the frame of speech under analysis are warped to the Mel-scale using a filterbank. Mel-scale warping is motivated by the findings of psychoacoustics that suggest that human perception of different frequency components is nonlinear. In other words, the use of Mel-filterbank is to mimic the human perception mechanism. The Mel-filterbank provides better resolution for low-frequency contents, while a greater degree of averaging happens in the high-frequency range. As a result, the spectral information present in the high-frequency region of speech is down-sampled by Mel-scale warping. Since the speaker characteristics are predominantly reflected in the high-frequency components [21, 40], the use of Mel-filterbank was observed to degrade the performance of automatic speaker recognition system. Motivated by this fact, the use of linear filterbank was proposed in [40]. On other hand, Mel-scale warping helps in the case of ASR since the speaker characteristics get suppressed to a large extent.

In this paper, the relative importance of linear and Mel-filterbanks in the context of ASR has been studied. For sake of completeness, the role of inverse-Mel-filterbank is also explored. This study is motivated by the fact that, in the case of children's speech, a significant amount of relevant spectral information is present in the higher-frequency region [3,6,25,30]. Therefore, wide-band speech data (sampled at 16 kHz rate) are preferred in the case of children's ASR. As mentioned earlier, the resolution of Mel-filterbank decreases as the frequency is increased. Hence, down-sampling the spectra is not beneficial in those cases where the speech data are from high-pitched child speakers. This is also true in the case of adult females as observed in this work. On the contrary, providing equal resolution to all the frequencies should be the preferred choice. In other words, using linear filterbank will be more effective when the speech data are from high-pitched speakers.

In order verify these claims, separate set of front-end features were extracted by applying Mel, inverse-Mel and linear filterbanks. Next, using each type of feature, separate ASR systems ware trained on adults' speech data from both male and female speakers. The ASR systems were evaluated using two different test sets. The first test set consisted of speech data from adult male and female speakers, while the second

one was comprised of data from children. To get better insight, the adults' speech test set was further split into two parts based on the gender of the speaker. The use of Mel-filterbank was noted to be more effective when the test speech data were from adult male speakers. In those cases when speech data were from adult females or children, employing linear or inverse-Mel-filterbank was observed to be more effective. In order to further boost confidence in those observations, linear frequency warping through vocal-tract length normalization (VTLN) [20] and explicit pitch scaling were also explored to suppress the ill effects induced by other speaker-dependent acoustic mismatch factors. In the context of children's speech recognition, use of VTLN [31,32] or explicit pitch scaling is reported to be very effective [1,15,16]. Even after the inclusion of VTLN and pitch scaling, the use of linear filterbank was noted to be more effective when the test data were from high-pitched speakers (adult females and children).

Some of the recent works have shown that pitch-induced acoustic mismatch also degrades the recognition when children's speech is transcribed using ASR system trained on adults' speech [11,12]. In the case of MFCC features, the pitch-induced spectral distortions arise due to insufficient smoothing of pitch harmonics despite the use of low-time liftering [13,38]. In such mismatched scenario, spectral smoothing is reported to be highly effective [10,37]. Sufficient spectral smoothing can be achieved by reducing the length of the low-time lifter as shown in [13] or by low-rank feature projection as explored in [34,36]. However, the recognition rates for the adults' speech got severely deteriorated due loss of relevant spectral information when cepstral truncation or low-rank projection was employed. Recently, a front-end speech parameterization technique exploiting TANDEM-STRAIGHT-based spectral smoothing was proposed for robust ASR [35]. The resulting front-end features, referred to as TS-MFCC, were reported to enhance the recognition performance not only for children's speech but also for adults'. Motivated by the effectiveness of TS-MFCC in ASR, the role of employing linear filterbank along with TANDEM-STRAIGHT-based spectral smoothing has also been studied in this paper. When linear filterbank is used in place of Mel-filterbank, the recognition performance is observed to improve with respect to high-pitch speakers.

The rest of this paper is organized as follows: In Sect. 2, motivation for studying the role of linear and inverse-Mel-filterbank in ASR is discussed. In Sect. 3, the experimental evaluations demonstrating the effectiveness of linear and inverse-Mel-filterbanks are presented. The effect of combining pitch-adaptive spectral estimation based on TANDEM-STRAIGHT and linear filterbank is studied in Sect. 4. Finally, the paper is concluded in Sect. 5.

2 Motivation for Studying the Role of Filterbanks in ASR

As mentioned earlier, the MFCC features are one of the most dominant front-end features in the context of ASR [39]. In order to make the following discussion more complete, we have first briefly described the MFCC feature extraction process once again. Given the raw speech data, the steps involved in the extraction of MFCC features are as follows [7]: Speech signal is first processed through a pre-emphasis filter in order

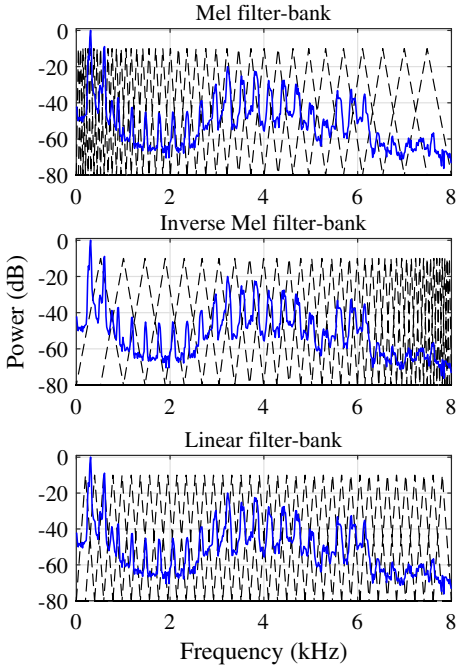

Fig. 1 Log-compressed power spectrum corresponding to the central portion of a voiced frame of speech from high-pitched child speaker. The 40-channel Mel-, inverse-Mel- and linear filterbanks are superimposed over the spectrum

to emphasize the higher-frequency components. Next, short-time frames of speech signal are created using overlapping Hamming windows. Typically, the duration of the analysis window is 20–30 ms with an overlap of 50%. This is followed by deriving the frequency domain representation for each of the short-time frames. Discrete Fourier transform (DFT) is used for this purpose. The phase information is discarded from the resulting short-term spectrum. The magnitude or the power spectrum is then warped to Mel-scale using a set of nonlinearly spaced filters. The Mel-filterbank is a set of triangular Mel-weighted filters. Next, logarithmic compression is performed followed by the application of discrete cosine transform (DCT) to derive a set of de-correlated cepstral coefficients. Finally, a low-time liftering operation is performed to discard the higher-order coefficients. In the context of ASR, only the first 13 coefficients are retained and they are collectively known as MFCC features.

The primary idea behind warping the linear frequencies to Mel-scale is to mimic the nonlinear behavior of human speech perception mechanism. The frequency resolution of the Mel-filterbank decreases as one moves toward the high-frequency region [40]. This fact is evident form the spectral plots shown in Fig. 1 (top pane). The short-time log-compressed power spectrum corresponding to the central portion of a voiced frame of speech is plotted. The 40-channel Mel-filterbank is superimposed over the spectrum. The speech data used for this analysis are from a high-pitched child speaker. It is to note that wide-band speech data are used for all the analyses presented in this paper. As clearly visible from the plots, the degree of averaging is more in the high-frequency region. This behavior of Mel-filterbank has an added advantage that the

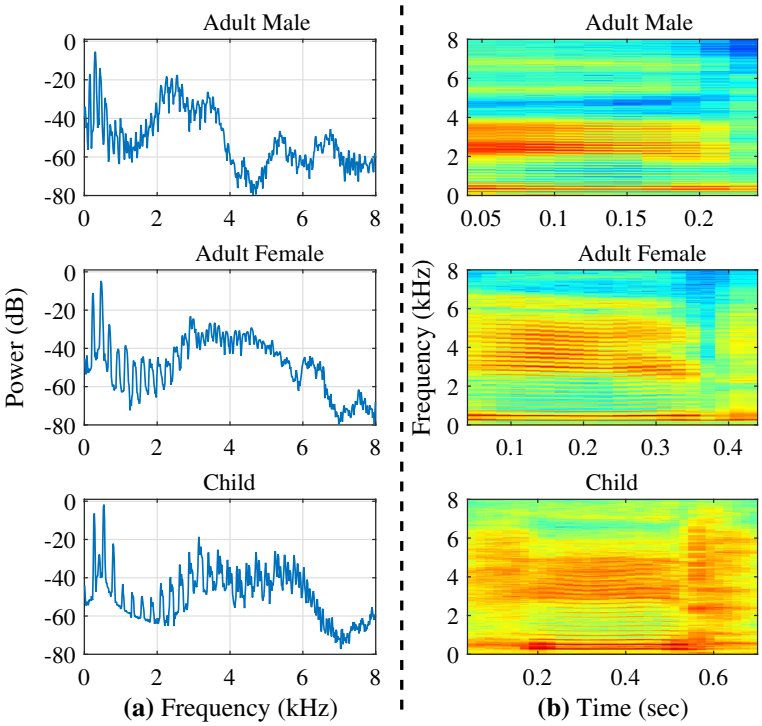

Fig. 2 **a** Power spectra for vowel/IY/extracted from the speech data belonging to adult male, adult female and child speakers, respectively. Short-time frames corresponding to the central portion of the same vowel were used for this analysis. **b** The corresponding spectrograms are shown

speaker-dependent acoustic attributes are smoothed out. This, in turn, is beneficial for ASR task where speaker independence is highly desired.

When dealing with children's speech or speech from high-pitched speakers like adult females, the down-sampling of spectral information in high-frequency components has a downside. As already stated, there is a significant amount of spectral information in the higher-frequency region that is important for ASR. Earlier works have shown that the formant frequencies are scaled up in the case of children's speech [9,19,29]. To demonstrate this characteristic of speech, the log-compressed power spectra corresponding to the central portion of vowel/IY/along with the corresponding spectrograms are plotted in Fig. 2. Scaling up of formant frequencies in the case speech data from adult female and child is easily noticeable. At the same time, the power is significantly high even in the 4–8 kHz frequency range. On the other hand, the power in 4–8 kHz frequency range is very less when the data are from adult male speaker. The spectral information in the high-frequency region should also be effectively preserved in order to improve the recognition performance with respect to

high-pitched speakers. Motivated by these observations and findings of earlier works on children's speech, the role of inverse-Mel and linear filterbanks are studied in this paper.

The relation between linear frequency scale (f in Hz) and Mel-frequency scale (m) is as follows:

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right). \tag{1}$$

Generally, each filter in the filterbank is triangular in shape and with a peak response equal to unity at the center frequency. The frequency response decreases linearly toward zero when it reaches the center frequencies of the adjacent filters as shown in Fig. 1 (top pane). The set of M filters can be designed using the following equation:

$$H_{\mathrm{m}}(k) = \begin{cases} 0, & k < f(m-1) \\ \frac{2(k-f(m-1))}{f(m)-f(m-1)}, & f(m-1) \le k < f(m) \\ 1, & k = f(m) \\ \frac{2(f(m+1)-k)}{f(m+1)-f(m)}, & f(m) < k \le f(m+1) \\ 0, & k > f(m+1) \end{cases} \tag{2}$$

with m ranging from 0 to $M-1$.

The inverse-Mel-scale is defined as the complement of Mel-scale [4,28]. The inverse-Mel-filterbank is obtained simply by flipping the original Mel-filterbank around the mid point, i.e., $f = 4$ kHz as shown in Fig. 1 (middle pane). Unlike Mel-filterbank, better resolution is obtained in the higher-frequency region. This is evident from the log-compressed power spectrum with inverse-Mel-filterbank superimposed over it which is shown in Fig. 1 (middle pane). The front-end features obtained by replacing the Mel-filterbank with inverse-Mel-filterbank are referred to as inverse-MFCC (IMFCC) in the remainder of this paper. The linear filterbank provides equal resolution to all the frequency components, and the same is evident from Fig. 1 (bottom pane). The front-end cepstral features obtained by using linear filterbank are called linear-frequency cepstral coefficients (LFCC) in this work. In the following section, we present the experimental evaluations demonstrating the relative effectiveness of MFCC, IMFCC and LFCC features in the context of ASR.

3 Experimental Evaluations

The simulation studies performed for evaluating the relative effectiveness of MFCC, IMFCC and LFCC features are presented in this section.

3.1 Experimental Setup

3.1.1 Speech Corpora

The speech data used for training the ASR system were obtained from the **British English** speech corpus WSJCAM0 [27]. The train set created from WSJCAM0 consisted of 15.5 h of speech data from 92 adult speakers (both male and female). The total number of utterances in the train set was 7852 with a total 132,778 words. In order to evaluate the effectiveness of the explored front-end features, three different test sets were created. The details of those test sets are as follows:

- *AD-Set* This test set was derived from the WSJCAM0 corpus and consisted of 0.6 h of speech from 20 adult male as well as female of speakers with a total of 5608 words.
- *ADF-Set* This test set was derived by splitting AD-Set and consisted of nearly 0.3 h of speech from 10 adult female speakers with a total of 2864 words.
- *CH-Set* For evaluating recognition performance with respect to children's speech, a test set derived from PF-STAR [2] **British English** speech database was employed. This test set consisted of 1.1 h of speech data from 60 child speakers with a total of 5067 words. The age of the child speakers in this test set was in between 4 and 14 years.

The experimental studies reported in this paper were performed on wide-band speech data (sampled at 16 kHz rate). The PF-STAR database is originally sampled at 22,050 samples per second, so down-sampling was done for consistency.

3.1.2 Front-End Feature Extraction

In order to extract the three kinds of front-end features, speech data were first high-pass filtered with pre-emphasis factor being 0.97. Short-time frames were then created using overlapping Hamming windows of length 20 ms with frame shift of 10 ms. For MFCC, IMFCC as well as LFCC, 40-channel filterbank was used to extract the 13-dimensional base features. Next, the base features were time-spliced by appending 4 frames to the left and to the right of the current analysis frame to it. The resulting 117-dimensional features vectors were then projected to 40 dimensional subspace using linear discriminant analysis (LDA) and maximum-likelihood linear transformation (MLLT) to derive the final feature vectors. This was followed by application of cepstral mean and variance normalization (CMVN) to all the front-end feature kinds. In addition to CMVN, feature normalization was also performed using feature-space maximum-likelihood linear regression (fMLLR) for boosting robustness toward speaker-dependent variations [26].

3.1.3 ASR System Architecture

The ASR systems were developed on the 15.5 h adults' speech data from the WSJCAM0 speech corpus. The Kaldi speech recognition toolkit [24] was used for ASR system development and evaluation. Context-dependent hidden Markov models

(HMM) were used in this work. Decision tree-based state tying was performed to fix the maximum number of tied-states (senones) at 2000. Observation densities for the HMM states were modeled using deep neural networks (DNN) [5,14]. A number of recent works have shown that acoustic modeling based on DNN-HMM framework can significantly improve children's speech recognition [22,23,31–33]. Prior to learning the DNN parameters, the fMLLR-normalized feature vectors were time-spliced once more considering a context size of 9 frames. The number of hidden layers in the DNN was chosen as 8. Each of the hidden layers consisted of 1024 hidden nodes with tan$h$ nonlinearity. The initial learning rate was selected as 0.015 which was reduced to 0.002 in 20 epochs. Extra 10 epochs were employed after reducing the learning rate. The minibatch size for neural net training was selected as 512. The initial state-level alignments employed in DNN training were generated using a Gaussian-mixture-model-based system.

While decoding adults' speech test, the standard MIT-Lincoln 5k Wall Street Journal bigram language model (LM) was used. The MIT-Lincoln LM has a perplexity of 95.3 with respect to adults' test set with no out-of-vocabulary (OOV) words. The employed lexicon consisted of 5850 words along with the pronunciation variations. While decoding the children's speech test, on the other hand, a 1.5 k bigram LM was used. This bigram LM was trained on the transcripts of speech data in PF-STAR after excluding the test set. A lexicon consisting of 1969 words including the pronunciation variations was used. The word error rate (WER) metric was used for evaluating the recognition performance.

3.2 Baseline System Performance

The baseline WERs for the adults' and children's speech test sets obtained by using MFCC features are presented in Table 1. On comparing the WERs for AD-Set and CH-Set, a huge difference is noted. One of the factors for the observed difference is that the Mel-scale warping leads to down-sampling of spectral information in high-frequency components as discussed earlier. Consequently, the use of IMFCC and LFCC improves the recognition performance with respect to children's speech as evident from the WERs enlisted in Table 1. At the same time, the recognition rates for adults' speech are noted to degrade when LFCC features are used. But the loss incurred in the case of adults' speech is much less when compared to the gain obtained for children's

Table 1 WERs for the adults' and children's speech test sets with respect to the acoustic models trained on adults' speech

Test set	WER (in %) for different acoustic features			Relative imp. over MFCC with LFCC (%)
	MFCC	IMFCC	LFCC	
AD-Set	5.87	5.93	6.11	− 4.1
CH-Set	19.37	18.14	16.35	15.6

The WERs are given for the cases when MFCC, IMFCC and LFCC features are used to train DNN-HMM-based systems

Table 2 WERs for adults', adult females' and children's speech test sets with respect to acoustic models trained on adults' speech

Acoustic feature	Test set	WER (in %)		
		Baseline	VTLN	Pitch scaling
MFCC	AD-Set	5.87	5.83	5.81
	ADF-Set	6.35	6.11	5.67
	CH-Set	19.37	17.00	13.11
IMFCC	AD-Set	5.97	5.95	5.93
	ADF-Set	6.10	5.93	5.28
	CH-Set	18.14	16.56	12.86
LFCC	AD-Set	6.11	6.10	6.06
	ADF-Set	5.94	5.84	**5.23**
	CH-Set	16.35	14.89	**12.19**

WERs are given for the cases when MFCC, IMFCC and LFCC features are used to train the DNN-HMM systems. The WERs are also tabulated for the cases when VTLN and explicit pitch scaling are employed for reducing the acoustic mismatch

speech. This fact is highlighted by the percentage of relative improvement in the MFCC obtained by using LFCC features given in the last column of Table 1.

It may be argued that, by including children's speech data in the train set, the differences will be reduced as reported in earlier works. In order to demonstrate that by folding sufficient amount of speech data into training the ill effects of down-sampling, the spectral information present in high-frequency region cannot be addressed, the adult test set was split into two parts based on the gender of the speaker. The test set created by taking speech data only from female speakers (ADF-Set) was then decoded using the adult data trained acoustic models. The WERs for this study are given in Table 2. The use of IMFCC and LFCC features is noted to reduce the WER significantly when compared to MFCC features. A relative improvement of 6.5% is obtained when LFCC is used instead of MFCC. It is to note that the training set derived from WSJCAM0 database contains a sufficient amount of speech data from adult female speakers as well [27]. Despite that, providing higher resolution to high-frequency components (IMFCC) or equal resolution to all the frequency components (LFCC) helps when the test speech is from adult female speakers.

It may also be argued that there are several other factors of acoustic mismatch that lead to degradation in the recognition performance, especially in the case of children's speech. In order to counter it, the role of VTLN and pitch scaling is explored to reduce the acoustic mismatch resulting from upscaling of fundamental and formant frequencies noted in the case of children' speech. These studies are presented in the following subsection.

3.3 Application of VTLN and Pitch Scaling

The vocal organs of children and adult females are smaller when compared to that of adult males [9,29]. As a consequence, formant frequencies are upscaled when

the speech data are either from adult females or children. Linear-frequency warping through VTLN is reported to address the ill effects of upscaling of formants [32, 38]. VTLN was implemented by extracting acoustic features after varying the linear frequency warping factor from 0.88 to 1.12 in steps of 0.02. The warped feature vectors were then forced-aligned against the acoustic model under the constraints of the first-pass hypothesis obtained by decoding the unwarped features. The set of features that resulted in highest likelihood were chosen to be optimal. The optimally warped feature vectors were then re-decoded after performing fMLLR-based feature normalization. The effect of concatenating VTLN and fMLLR on the MFCC, IMFCC and LFCC features is demonstrated using WERs given in Table 2. Large reductions in WER are noted by the application of VTLN in the case of children's speech. The observed reductions in the case of adult females is not that large. At the same time, the LFCC features are still noted to be superior to MFCC for both ADF-Set and CH-Set test sets.

Apart from formant scaling, even the fundamental frequency or pitch is noted to change due differences in vocal-tract geometry. Consequently, the fundamental frequency is observed to be higher in the case of children as well as adult female speakers. Pitch-induced acoustic mismatch severely degrades the ASR performance as reported in [33,38]. The ill effects of pitch variations can be compensated by explicit pitch modification as reported in [16]. Motivated by that work, pitch modification was also explored in order to improve the recognition performance with respect to high-pitched speakers. The pitch scaling technique reported in [1] was explored for this purpose. The tunable pitch compensation factor (semitone) was varied from -12 to 12 in steps of 1 to vary the pitch of the speech data being analyzed. The optimal compensation factor was chosen via a maximum-likelihood grid search described earlier. The WERs obtained by suitably modifying the pitch are given in Table 2. Similar to the case of VTLN, large reductions in WER are observed in the case of children's speech. Even for adult females, the reductions in WER are significant. The use of LFCC features is noted to be superior in this case as well. The reduction in WER is larger in the case of pitch scaling than that obtained with the application VTLN. Pitch scaling is performed by re-sampling the speech data followed by time-scale modification. Re-sampling results in rescaling of the formant frequencies as well. Consequently, VTLN is done in a implicit manner when explicit pitch modification is performed.

3.4 Experiments Employing Filterbank Features

In recent works on ASR, the log-compressed outputs of filterbank have been employed as front-end acoustic feature with DNN-HMM-based system [14]. Motivated by those studies, we have also explored filterbank outputs as acoustic features. The filterbank features obtained by Mel, inverse-Mel and linear filterbanks are referred to as MFBANK, IMFBANK and LFBANK features, respectively, in this work. The WERs obtained by using filterbank features are given in Table 3. Reduced WERs are obtained when inverse-Mel or linear filterbank is used in this case as well.

Table 3 WERs for adult females' and children's speech test sets with respect to acoustic models trained on adults' speech

Acoustic feature	Test set	WER (in %)		
		Baseline	VTLN	Pitch scaling
MFBANK	ADF-Set	6.20	6.05	5.53
	CH-Set	18.29	16.35	12.82
IMFBANK	ADF-Set	5.95	5.45	5.31
	CH-Set	17.45	16.14	12.52
LFBANK	ADF-Set	5.42	5.29	**5.04**
	CH-Set	15.68	14.43	**11.98**

WERs are given for the cases when MFBANK, IMFBANK and LFBANK features are used to train the DNN-HMM systems. The WERs are also tabulated for the cases when VTLN and explicit pitch scaling are employed for reducing the acoustic mismatch

Table 4 WERs for the adults' and children's speech test sets with respect to adult data trained DNN system demonstrating the effectiveness of TS-MFCC over MFCC and LFCC

Test set	WER (in %)		
	MFCC	TS-MFCC	LFCC
AD-Set	5.87	5.28	6.11
CH-Set	19.37	15.85	16.35

4 Combining Linear Filterbank with Pitch-Adaptive Spectral Estimation

As highlighted earlier, the conventional approach for extracting front-end acoustic features does not explicitly depend on pitch-adaptive signal processing. This leads to insufficient smoothing of the spectra, especially for the high-pitched speakers. Spectral smoothing is essential in order to reduce the ill effects of pitch harmonics. Kawahara et al. proposed a pitch-adaptive spectral analysis technique which was named as STRAIGHT [18]. The MFCC features derived using the STRAIGHT-based spectra were employed for ASR in [8] and were not found to be very effective. This was mainly due to a smoothing function used after the pitch-adaptive windowing which led to over-smoothing. Further, legacy STRAIGHT is reported to be computationally expensive. To alleviate these problems, TANDEM STRAIGHT was introduced for spectrum estimation [17].

The role of pitch-adaptive estimation via TANDEM STRAIGHT in ASR was studied in [35]. The resulting front-end features (TS-MFCC) were reported to be better than the existing MFCC. This fact is re-verified by the WERs given in Tables 4. The WERs obtained by employing LFCC features are also enlisted for proper contrast. Its evident from the WERs given in Tables 4 that TS-MFCC features outperform MFCC as well as LFCC features for both adults' and children's speech test sets. Yet, compared to MFCC, the WERs obtained by the use of LFCC features are much closer to those obtained through TS-MFCC.

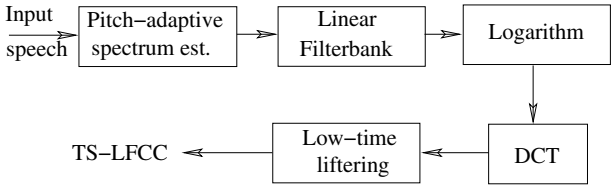

Fig. 3 Block diagram outlining the proposed front-end speech parameterization technique

Table 5 WERs for adult, adult females' and children's speech tests set with respect to adult data trained DNN system demonstrating the effect of using linear filterbank in place of Mel-filterbank along with pitch-adaptive spectrum estimation

Test set	WER (in %)		% Relative improvement
	TS-MFCC	TS-LFCC	
AD-Set	5.28	5.46	− 3.1
ADF-Set	5.96	5.78	3.2
CH-Set	15.85	13.72	13.44

Motivated by the success of TANDEM-STRAIGHT-based spectral smoothing, the effect of including linear filterbank instead of Mel-filterbank was explored next. The overall process of extracting the proposed features employing pitch-adaptive spectral estimation and linear filterbank is summarized in Fig. 3. The proposed features are, therefore, referred to as TANDEM-STRAIGHT linear filterbank cepstral coefficients (TS-LFCC) in this paper. The introduced modifications provide added robustness toward speaker-dependent acoustic variations. The same is experimentally validated in the following.

4.1 Experimental Results

The effect of including linear filterbank along with TANDEM-STRAIGHT-based spectral smoothing is demonstrated by the WERs enlisted in Table 5. Additive reductions in WER are obtained for adult females' and children's speech when linear filterbank is used instead of Mel-filterbank. On the other hand, a slight degradation is noted in the case of adults' speech test set. These results further establish the fact that the use of linear filterbank is more beneficial when the speech data are from high-pitched speakers. On comparing the MFFC-based DNN baseline for children's speech (19.37%), a relative improvement of 29.17% is obtained when the proposed TS-LFCC features are used.

4.2 Inclusion of VTLN and Pitch Scaling

The effectiveness of performing VTLN as well as explicit pitch scaling was explored next. The WERs for those studies are given in Table 6. As noted in the case of MFCC/LFCC, both VTLN and pitch scaling are highly effective when combined

Table 6 WERs for adults', adult females' and children's speech test sets with respect to adult data trained DNN systems demonstrating the effect of combining VTLN or pitch scaling with TS-MFCC and TS-LFCC features

Acoustic feature	Test set	WER (in %)		
		Baseline	VTLN	Pitch scaling
TS-MFCC	AD-Set	5.28	5.31	5.39
	ADF-Set	5.96	5.85	5.54
	CH-Set	15.85	14.37	11.67
TS-LFCC	AD-Set	5.46	5.51	5.45
	ADF-Set	5.78	5.66	**5.32**
	CH-Set	13.72	12.87	**10.92**

Table 7 WERs for adults', adult females' and children's speech test sets with respect to adult data trained DNN systems demonstrating the effect of combining VTLN or pitch scaling with TS-MFBANK and TS-LFBANK features

Acoustic feature	Test set	WER (in %)		
		Baseline	VTLN	Pitch scaling
TS-MFBANK	AD-Set	5.24	5.29	5.33
	ADF-Set	5.73	5.61	5.47
	CH-Set	15.15	13.94	11.28
TS-LFBANK	AD-Set	5.39	5.47	5.45
	ADF-Set	5.64	5.51	**5.32**
	CH-Set	13.28	12.17	**10.56**

with TS-MFCC or TS-LFCC features. At the same time, the WERs obtained by using TS-LFCC features are significantly better when the speech data are from high-pitched speakers.

In this work, the effect of pitch-adaptive spectral smoothing on filterbank features has also been studied. The resulting features are referred to as TS-MFBANK when Mel-filterbank is used while TS-LFBANK when linear filterbank is employed . The WER obtained by employing TS-MFBANK and TS-LFBANK features is given in Table 7. Compared to the baseline WER obtained by employing MFBANK features (18.29%, see Table 3), significant reductions in WERs are observed with the application of TANDEM-STRAIGHT-based spectral smoothing. Applying pitch scaling leads to further reductions in WER when the speech data are from high-pitched speakers. At the same time, the use of linear filterbank is noted to consistently yield better recognition performances.

5 Conclusion

In this paper, the role Mel-, inverse-Mel- and linear filterbanks are studied in the context of ASR task. The presented work is motivated by the fact that there is a significant

amount of relevant spectral information present in the high-frequency region when the speech data are from adult female and child speakers. Consequently, down-sampling the spectral information in that range through Mel-filterbank reduces the recognition performance. The inverse-Mel and linear filterbanks provide better resolution to the high-frequency components. Therefore, significant improvements are noted when IMFCC or LFCC features are used when the speech data being transcribed are from adult female or child speakers. In order to further boost our confidence in the observed improvements, the role of VTLN and explicit pitch scaling has also been explored. Even after the application of VTLN or pitch scaling, LFCC features are noted to be better than MFCC features. In addition to that, the effect of combining pitch-adaptive spectral estimation with linear filterbank has also been explored. Added improvements in recognition performance are noted with the inclusion of pitch-adaptive spectral estimation.

References

1. W. Ahmad, S. Shahnawazuddin, H.K. Kathania, G. Pradhan, A.B. Samaddar, Improving children's speech recognition through explicit pitch scaling based on iterative spectrogram inversion. in *Proceedings of INTERSPEECH* (2017)
2. A. Batliner, M. Blomberg, S.D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, M. Wong, The PF_STAR children's speech corpus. in *Proceedings of INTERSPEECH*, pp. 2761–2764 (2005)
3. D. Byrd, S. Yildirim, S. Narayanan, S. Khurana, Acoustic analysis of preschool children's speech. in *Proceedings of 15th ICPhS Barcelona*, pp. 949–952 (2003)
4. S. Chakroborty, A. Roy, G. Saha, Improved closed set text-independent speaker identification by combining MFCC with evidence from flipped filter banks. Int. J. Electr. Comput. Energ. Electron. Commun. Eng. **2**(11), 2554–2561 (2008)
5. G. Dahl, D. Yu, L. Deng, A. Acero, Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. IEEE Trans. Speech Audio Process. **20**(1), 30–42 (2012)
6. S. D'Arcy, M. Russell, A comparison of human and computer recognition accuracy for children's speech. in Proceedings of INTERSPEECH, pp. 2187–2200 (2005)
7. S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. **28**(4), 357–366 (1980)
8. G. Garau, S. Renals, Combining spectral representations for large-vocabulary continuous speech recognition. IEEE Trans. Speech Audio Process. **16**(3), 508–518 (2008)
9. M. Gerosa, D. Giuliani, S. Narayanan, A. Potamianos, A review of ASR technologies for children's speech. in *Proceedings of Workshop on Child, Computer and Interaction*, pp. 7:1–7:8 (2009)
10. S. Ghai, Addressing pitch mismatch for children's automatic speech recognition. Ph.D. thesis, Department of EEE, Indian Institute of Technology Guwahati, India, 2011
11. S. Ghai, R. Sinha, A study on the effect of pitch on LPCC and PLPC features for children's ASR in comparison to MFCC. in *Proceedings of INTERSPEECH*, pp. 2589–2592 (2011)
12. S. Ghai, R. Sinha, Analyzing pitch robustness of PMVDR and MFCC features for children's speech recognition. in *Proceedings of Signal Processing and Communications (SPCOM)* (2010)
13. S. Ghai, R. Sinha, Exploring the role of spectral smoothing in context of children's speech recognition. in *Proceedings of INTERSPEECH*, pp. 1607–1610 (2009)
14. G.E. Hinton, L. Deng, D. Yu, G. Dahl, A.R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, B. Kingsbury, Deep neural networks for acoustic modeling in speech recognition. Signal Process. Mag. **29**(6), 82–97 (2012)
15. H.K. Kathania, S. Shahnawazuddin, N. Adiga, W. Ahmad, Role of prosodic features on children's speech recognition. in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5519–5523 (2018)

16. H.K. Kathania, W. Ahmad, S. Shahnawazuddin, A.B. Samaddar, Explicit pitch mapping for improved children's speech recognition. Circuits Syst. Signal Process. **37**(5), 2021–2044 (2017)
17. H. Kawahara, M. Morise, Technical foundations of TANDEM-STRAIGHTs, a speech analysis, modification and synthesis framework. Sadhana **36**(5), 713–727 (2011)
18. H. Kawahara, I. Masuda-Katsuse, A. De Cheveigné, Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds. Speech Commun. **27**(3), 187–207 (1999)
19. R.D. Kent, Anatomical and neuromuscular maturation of the speech mechanism: evidence from acoustic studies. JHSR **9**, 421–447 (1976)
20. L. Lee, R. Rose, A frequency warping approach to speaker normalization. IEEE Trans. Speech Audio Process. **6**(1), 49–60 (1998)
21. H. Lei, E. Gonzalo, Mel, linear, and antimel frequency cepstral coefficients in broad phonetic regions for telephone speaker recognition. in *Proceedings of INTERSPEECH*, pp. 2323–2326 (2009)
22. H. Liao, G. Pundak, O. Siohan, M.K. Carroll, N. Coccaro, Q. Jiang, T.N. Sainath, A.W. Senior, F. Beaufays, M. Bacchiani, Large vocabulary automatic speech recognition for children. in *Proceedings of INTERSPEECH*, pp. 1611–1615 (2015)
23. A. Metallinou, J. Cheng, Using deep neural networks to improve proficiency assessment for children English language learners. in *Proceedings of INTERSPEECH*, pp. 1468–1472 (2014)
24. D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, K. Vesely, The Kaldi speech recognition toolkit. in *Proceedings of ASRU* (2011)
25. M.R. Qun Li, An analysis of the causes of increased error rates in children's speech recognition. in *Proceedings of ICSLP2002*, Sept 2002
26. S.P. Rath, D. Povey, K. Veselý, J. Černocký, Improved feature processing for deep neural networks. in *Proceedings of INTERSPEECH* (2013)
27. T. Robinson, J. Fransen, D. Pye, J. Foote, S. Renals, WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition. in *Proceedings of ICASSP*, vol. 1, pp. 81–84 (1995)
28. A. Roy, G. Saha, S. Majumdar, S. Chakroborty, Capturing complementary information via reversed filter bank and parallel implementation with MFCC for improved text-independent speaker identification. in *Proceedings of International Conference on Computing: Theory and Applications(ICCTA)*, pp. 463–467 (2007)
29. M. Russell, S. D'Arcy, Challenges for computer recognition of children's speech. in *Proceedings of Speech and Language Technologies in Education (SLaTE)* (2007)
30. M. Russell, S. D'Arcy, L. Qun, The effects of bandwidth reduction on human and computer recognition of children's speech. IEEE Signal Process. Lett. **14**(12), 1044–1046 (2007)
31. R. Serizel, D. Giuliani, Vocal tract length normalisation approaches to DNN-based children's and adults' speech recognition. in *Proceedings of Spoken Language Technology Workshop (SLT)*, pp. 135–140 (2014)
32. R. Serizel, D. Giuliani, Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children. Nat. Lang. Eng. **23**(3), 325–350 (2016)
33. S. Shahnawazuddin, A. Dey, R. Sinha, Pitch-adaptive front-end features for robust children's ASR. in *Proceedings of INTERSPEECH* (2016)
34. S. Shahnawazuddin, H. Kathania, R. Sinha, Enhancing the recognition of children's speech on acoustically mismatched ASR system. in *Proceedings of TENCON* (2015)
35. S. Shahnawazuddin, N. Adiga, H.K. Kathania, G. Pradhan, R. Sinha, Studying the role of pitch-adaptive spectral estimation and speaking-rate normalization in automatic speech recognition. Digit. Signal Process. **79**, 142–151 (2018)
36. S. Shahnawazuddin, H.K. Kathania, A. Dey, R. Sinha, Improving children's mismatched asr using structured low-rank feature projection. Speech Commun. **105**, 103–113 (2018)
37. R. Sinha, S. Ghai, On the use of pitch normalization for improving children's speech recognition. in *Proceedings of INTERSPEECH*, pp. 568–571 (2009)
38. R. Sinha, S. Shahnawazuddin, Assessment of pitch-adaptive front-end signal processing for children's speech recognition. Comput. Speech Lang. **48**, 103–121 (2018)
39. R. Vergin, D. O'Shaughnessy, A. Farhat, Generalized Mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition. IEEE Trans. ASSP **7**(5), 525–532 (1999)

40. X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson, S. Shamma, Linear versus Mel frequency cepstral coefficients for speaker recognition. in *Proceedings of ASRU*, pp. 559–564 (2011)

Affiliations

Hemant Kumar Kathania[1] · S. Shahnawazuddin[2] · Waquar Ahmad[3] · Nagaraj Adiga[4]

Hemant Kumar Kathania
hemant.ece@nitsikkim.ac.in

Waquar Ahmad
waquar@nitc.ac.in

Nagaraj Adiga
nagaraj@csd.uoc.gr

1 Department of ECE, National Institute of Technology Sikkim, Sikkim, India

2 Department of ECE, National Institute of Technology Patna, Patna, India

3 Department of ECE, National Institute of Technology Calicut, Calicut, India

4 Department of Computer Science, University of Crete, Rethymnon, Greece