Circuits, Systems, and Signal Processing (2024) 43:3020–3041
https://doi.org/10.1007/s00034-023-02592-z

Enhancing Children's Short Utterance-Based ASV Using Inverse Gamma-tone Filtered Cepstral coefficients

Shahid Aziz[1] · **S. Shahnawazuddin**[1]

Received: 13 April 2023 / Revised: 21 December 2023 / Accepted: 21 December 2023 /
Published online: 10 February 2024

Abstract

The task of developing an automatic speaker verification (ASV) system for children's speech is extremely challenging due to the dearth of domain-specific data. The challenges are further exacerbated in the case of short utterances of speech, a relatively unexplored domain in the case of children's ASV. Voice-based biometric systems require an adequate amount of speech data for enrollment and verification; otherwise, the performance considerably degrades. It is for this reason that the trade-off between convenience and security is gruelling to maintain in practical scenarios. In this paper, we have focused on data paucity and preservation of the higher-frequency contents in order to enhance the performance of a short utterance-based children's speaker verification system. To deal with data scarcity, an out-of-domain data augmentation approach has been proposed. Since the out-of-domain data used are from adult speakers which are acoustically very different from children's speech, we have made use of techniques like prosody modification, formant modification, and voice conversion in order to render it acoustically similar to children's speech prior to augmentation. This helps in not only increasing the amount of training data but also in effectively capturing the missing target attributes which helps in boosting the verification. Further to that, we have resorted to concatenation of the classical Mel-frequency cepstral coefficients (MFCC) features with the Gamma-tone frequency cepstral coefficient (GTF-CC) or with the Inverse Gamma-tone frequency cepstral coefficient (IGTF-CC) features. The feature concatenation of MFCC and IGTF-CC is employed with the sole intention of effectively modeling the human auditory system along with the preservation of higher-frequency contents in the children's speech data. This feature concatenation approach, when combined with data augmentation, helps in further improvement in

✉ Shahid Aziz
shahida.phd20.ec@nitp.ac.in

S. Shahnawazuddin
s.syed@nitp.ac.in

[1] Department of Electronics and Communication Engineering, National Institute of Technology Patna, Patna, India

the verification performance. The experimental results testify our claims, wherein we have achieved an overall relative reduction of 38.5% for equal error rate.

Keywords Automatic speaker verification · Out-of-domain data augmentation · Gamma-tone frequency cepstral coefficient · Inverse Gamma-tone frequency cepstral coefficient

1 Introduction

Automatic speaker verification (ASV) is a biometric task of verifying the claimed identity of a speaker. It uses the characteristics of human voice/speech for authentication of the claimant. If the test speech sample, given at the time of verification, is closer to the target model (template), then the ASV system pronounces the claim to be genuine, else the speaker is declared an impostor. As compared to the other competing biometric technologies, voice biometrics stands strong due to its prompt, hassle-free, and error-free authentication. It revamps the surveillance by minimizing security breaches caused by compromised or stolen passwords, phishing, fraudsters, etc. Needless to say, these cyber scams and frauds can play havoc with anyone accessing online application tools ignorantly. Voice biometrics enables the system to spend less time in authenticating users and resetting passwords. ASV technology provides a low-cost biometric solution [18] and is, thus, increasingly gaining acceptance in remote access to applications including but not limited to banking and financial services, websites and networks, telephone and internet transaction authentication, audio signatures for digital documents, hands-free mobile authentication, authentication during a customer support call, biometric login, payment gateways, merchandising, forensics, healthcare and mobile workforces, social networking websites, e-games and e-learning tools. With the ever-growing need for surveillance and secured systems, ASV systems are destined to be ubiquitously present and a provisioner of a much-needed security shield to adults and children alike. Dismally, a major chunk of the works reported in the literature deals with the task of building an ASV system for the adult population. The literary works reported on building an ASV system for the more vulnerable lot, i.e., children, are regrettably unsubstantial [25, 28, 32].

State-of-the-art ASV systems are found to be very effective, incurring minimal error when they are fed with an adequately **larger** amount and **longer** duration of speech data. Apprehensively, most of the children's speech corpora are not easily available. Moreover, these are available in only a handful of languages spoken across the globe and limited in terms of hours of data. For the languages in which children's speech corpus is unavailable (zero-resource condition), developing an ASV system is quite a formidable task. Even if a limited amount of children's speech data is on offer (low-resource condition), developing an effective children's ASV system employing deep learning architectures is still very challenging. State-of-the-art ASV systems incorporate deep learning architectures that require estimating a huge number of parameters. This, in turn, requires a large amount of domain-specific data. To overcome with the issue arising with the low- and zero-resource conditions, a few earlier works on children's ASV have studied the impact of synthetically generating

speech data and then pooling it into training. Out-of-domain data augmentation has been reported to be effective in this regard [28].

An ASV system, in real-world applications, is also marred by constrained duration of speech utterances. Though this requirement can be fulfilled during training phase by some data augmentation techniques, it is not feasible to do the same during the testing phase. In forensics applications for instance, the employed ASV system is less likely to get sufficient data even for enrollment. In access control type cases, average utterance length is restricted to a few seconds only [17]. To the best of the authors' knowledge, there is hardly any work reported on children's ASV task using short utterances. Motivated by this gap in the research arena, the role of out-of-data augmentation techniques in the context of short utterance-based children's ASV task is explored in this paper. The effect of synthetically generating speech data from the available adults' speech corpus, which is acoustically similar to that of children's speech prior to augmentation, is analyzed in this study. The techniques used to address the dearth of domain-specific data explored in this paper include (i) voice conversion (VC) of adults' speech data using cycle-consistent generative adversarial network (C-GAN) [10], (ii) prosody modification (PM) [26, 27] of adults' speech, and (iii) up-scaling the formant frequencies (FM) [11, 14] of adults' speech data. All the explored techniques modify the attributes of adults' speech in order to render it acoustically similar to children's speech. The explored out-of-domain data augmentation technique is observed to be very effective as demonstrated through the experimental studies presented in this paper.

In addition to data augmentation, the effectiveness of frame-level concatenation of the most popular front-end acoustic feature, namely the MFCC with the GTF-CC or with the IGTF-CC, is also examined in this paper. In general, the MFCC are the most commonly used front-end acoustic features, which can capture speaker-specific characteristics. However, ASV systems based solely on MFCC features show susceptibility in a number of scenarios. Firstly, the performance of MFCC-based systems degrades drastically in the presence of background noise [6]. Secondly, the mel-scale in the standard MFCC is not the optimal auditory model [6]. Lastly, since the resolution of Mel-filter-bank decreases as the frequency is increased, the performance of MFCC-based ASV systems degrades in case of high-pitched speakers. The aforementioned facts have fuelled the authors' of this paper to delve into the role of another well-acclaimed front-end speech parameterization technique, namely the GTF-CC. Prior literary works have demonstrated that GTF-CC performs robustly in speaker verification tasks in the presence of additive noise over a wide range of signal-to-noise ratios [6]. Further, Gamma-tone filter-banks employed in the extraction of GTF-CC features exploit the advantages of human auditory system [22] as it is more subtle and similar to human auditory model. The Gamma-tone filter-bank is very similar to the rounded exponential function used in representing the magnitude response of the human auditory filters [8]. The Mel-filter-bank on the other hand is designed to model the human pitch perception mechanism [9]. The feature fusion model of MFCC and GTF-CC is thus expected to enhance the ASV performance. The authors in this backdrop have also explored a variant of the GTF-CC, namely the IGTF-CC. In this case, the employed filter-bank is obtained simply by flipping the Gamma-tone filter-bank around the midpoint. The Inverse Gamma-tone filter-bank is thus supposed to have better frequency resolution in higher-frequency range. The lower-frequency components

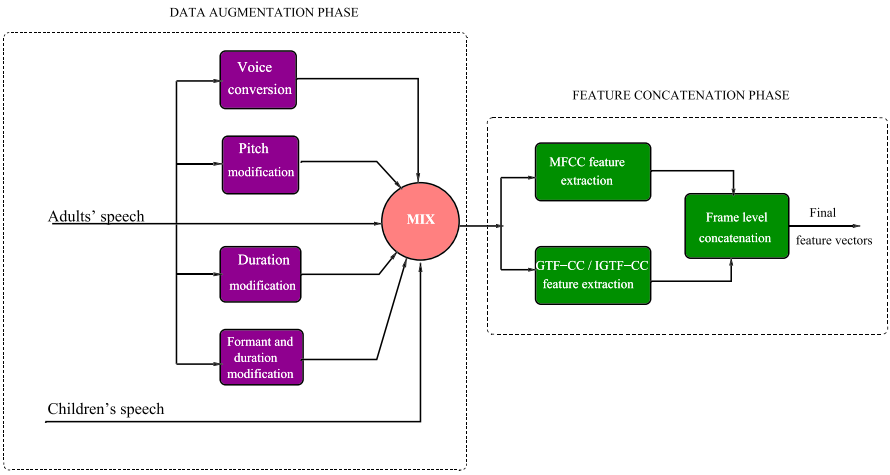

Fig. 1 Block diagram outlining the data augmentation and feature concatenation approaches proposed in this work in order to enhance the verification performance of a short utterance-based children ASV system

are down-sampled. As already highlighted, the use of Mel-filter-bank down-samples the spectral information in the higher-frequency range. The IGTF-CC due to its complementary nature of filter-bank is supposed to better capture the acoustic information in children's speech, which are otherwise averaged out by the MFCC features. The feature fusion model of MFCC and IGTF-CC is thus expected to outperform the traditional MFCC, leading to an enhanced children's ASV system. It is worth mentioning here to the best of the authors' knowledge that the role of IGTF-CC has not yet been explored in the context of children's speaker verification.

The aforementioned proposal of feature concatenation in addition to data augmentation is outlined in Fig. 1 and well validated in the experimental results section of the paper. The paper also illustrates the age group-wise as well as a gender-wise analysis of the children ASV performance to unravel the effect of data augmentation and feature concatenation. The proposed approach aids in diminishing the Equal Error Rate (EER) and Detection Cost Function (DCF) considerably as opposed to our baseline system trained exclusively on children's speech using MFCC features. The ASV systems for children's speech developed in this work for experimental evaluations employ x-vector-based speaker representation along with probabilistic linear discriminant analysis (PLDA)-based scoring.

The noteworthy contributions of this study are delineated as follows:

- A comprehensive examination of the children's speaker verification task using short utterances under low-resource conditions. As emphasized, there is a notable scarcity of research addressing the children's ASV task centered on short utterances;
- The efficacy of the suggested data augmentation strategy in mitigating the adverse impact resulting from the scarcity of domain-specific data is illustrated;

– The significance of the proposed frame-level concatenation of front-end acoustic features in preserving higher-frequency contents within children's speech data is emphasized and substantiated;
– An exhaustive examination of the children's ASV system is conducted, categorizing subjects by age group and gender, to assess the cumulative effects of data augmentation and feature concatenation.

The rest of this paper is organized as follows: Sect. 2 describes the proposed out-of-domain data augmentation techniques to deal with the scarcity of domain-specific data. In Sect. 3, we have talked about the authors' motivation to look beyond the traditional Mel-based filter-bank and delve into the scope of feature concatenation for children's ASV system. The experimental evaluations exhibiting the efficacy of our proposed technique are presented in Sect. 4. Eventually, conclusion and the future scope of the research work done in this paper are mentioned in Sect. 5.

2 Proposed Out-of-Domain Data Augmentation Technique

The state-of-the-art ASV system makes use of *x*-vectors-based speaker representation. For extracting *x*-vectors, a time-delay neural network (TDNN) [16, 30, 31] comprising a large number of hidden layers and hidden nodes per layer is trained. As already mentioned, one of the hurdles in the development of a reliable ASV system for children is the scarcity of domain-specific data. Hence, training an *x*-vector extractor on a limited amount of children's speech will result in sub-optimal performance. Out-of-domain data augmentation techniques can help mitigate this obstacle. However, it is worth highlighting here that the augmented data must have attributes similar to those of children's speech. Otherwise, the trained ASV system would fail to generalize well for unseen child speakers. Driven by this rationale, we present an out-of-domain data augmentation technique which is observed to be very effective in the context of children's ASV task using short utterances.

The proposed out-of-domain data augmentation technique is pictorially summarized in Fig. 1. In our approach, we use a limited amount of adults' speech for augmentation. As already mentioned, we have employed different techniques by which the acoustic attributes of adults' speech can be suitably modified and those are briefly discussed in the following:

In the first technique, the adults' speech was subjected to voice conversion (VC) using a cycle-consistent generative adversarial network (C-GAN) [10]. Nearly 10 min of speech data from each speaker group (adult and child speakers) was used to train the C-GAN. As a result of VC, adults' speech utterances sound very similar to children's speech as noted during the listening tests. Therefore, on pooling the voice-converted data, the issues of acoustic mismatch reduce to a large extent.

In the second technique, adults' speech was subjected to prosody modification prior to augmentation. It is well known that the pitch for children's speech is higher while the speaking rate is slower [14, 24]. Therefore, pitch of the speech data from the adult speakers was increased by a factor of 1.35 while the duration was increased by a factor of 1.4. These scaling factors were determined from earlier reported works on

children's speech recognition [26]. In order to perform prosody modification (PM), the technique reported in [21] was used. Again, pooling prosody-modified data ensure that the acoustic mismatch remains in check.
In the case of child speakers, the formant frequencies are higher as compared to adult speakers [14, 24]. Hence, in the third technique, the formant frequencies (FM) of adults' speech data were up-scaled by a factor of 0.08. At the same time, the speaking rate of adults' speech data was decreased by a factor of 1.4 through time-scale modification (TSM) [21]. This was done to compensate for the differences in speaking rates as discussed earlier. The mentioned scaling factors were adopted from the earlier works [13, 26]. Like in the case of VC and PM, pooling formant modified adults' data help in increasing the amount of training data while keeping acoustic mismatch in check to a large extent.

All the modified versions of adults' data are then pooled into training along with children's speech as well as the original adults' data. Consequently, the amount of training data is increased leading to a more robust estimation of model parameters. Moreover, modifying the acoustic attributes ensures that the developed ASV system does not get biased toward adult speakers. It is worth mentioning here that even though the aforementioned techniques of synthetically generating speech data are well acclaimed in literary works, their combined effectiveness in the context of children's ASV systems for short utterances is relatively uncharted.

3 Exploring the Role of Different Acoustic Features in Children ASV

3.1 Prior Art

As mentioned earlier, the MFCC features are one of the most popular and commonly used front-end acoustic features in the context of an ASV system. It is the first feature among the three front-end features explored in this paper. The step-wise process of extracting MFCC features is briefly described as follows:

- The speech signal is first high-pass filtered through a pre-emphasis filter in order to emphasize the higher-frequency components;
- Next, each of the speech utterances is analyzed into short-time frames using overlapping Hamming windows, followed by the computation of short-time Fourier transform (STFT);
- Spectral warping is then carried out using a set of non-linearly spaced filters, called Melody(Mel)-filter-bank. Mel-filter-bank is a set of triangular Mel-weighted filters as depicted in the top panel of Fig. 2;
- Logarithmic compression of the filtered power spectrum is then performed;
- The decorrelated real cepstrum (RC) is then obtained by applying discrete cosine transform (DCT);
- Finally, by low-time liftering of the real cepstrum, MFCC features are extracted which will eventually be fed as input for training any classifier.

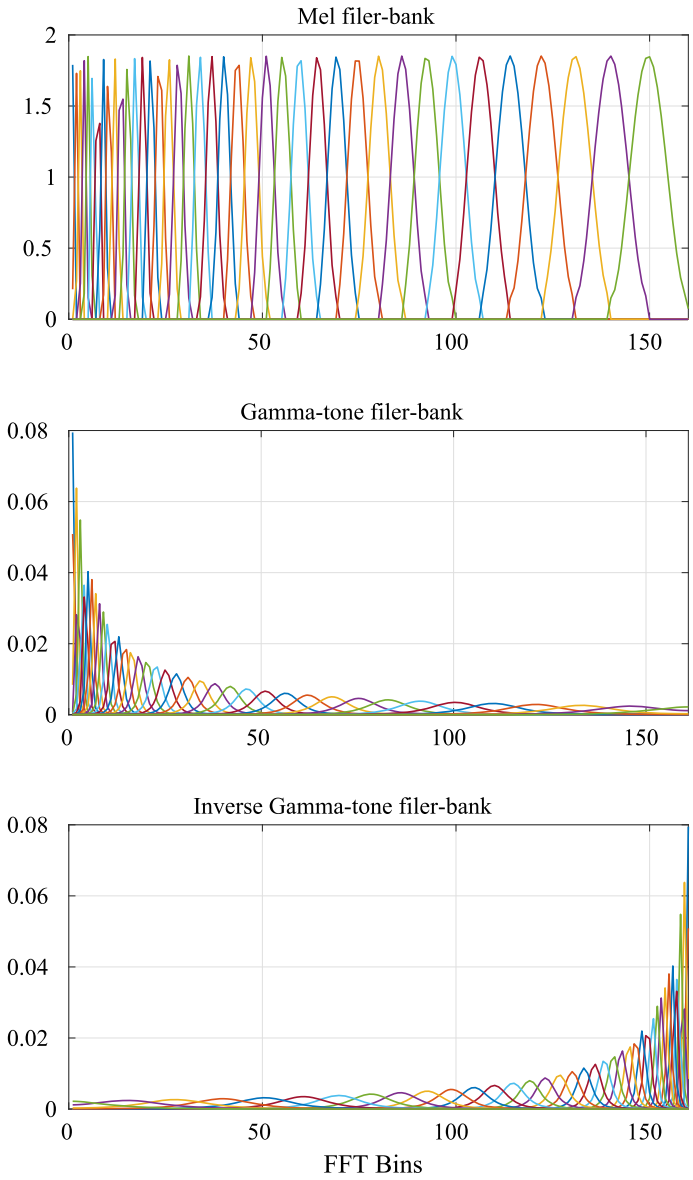

Fig. 2 Configuration of a Mel-, Gamma-tone and Inverse Gamma-tone filter-banks

3.2 Motivation for Exploring the Role of Feature Concatenation in Children ASV

MFCC features are the most conventional front-end acoustic features and have been the state of the art ever since its inception. They provide a compact and stable representation of the vocal tract of a speaker, significantly reducing the computational cost. The limitations of the MFCC features discussed in the previous section provide the

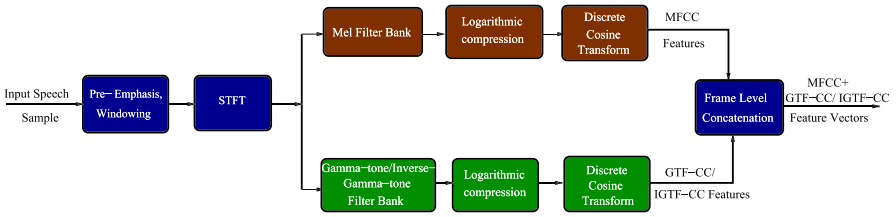

Fig. 3 Block diagram outlining the process of extracting concatenated MFCC features with either GTF-CC or IGTF-CC features

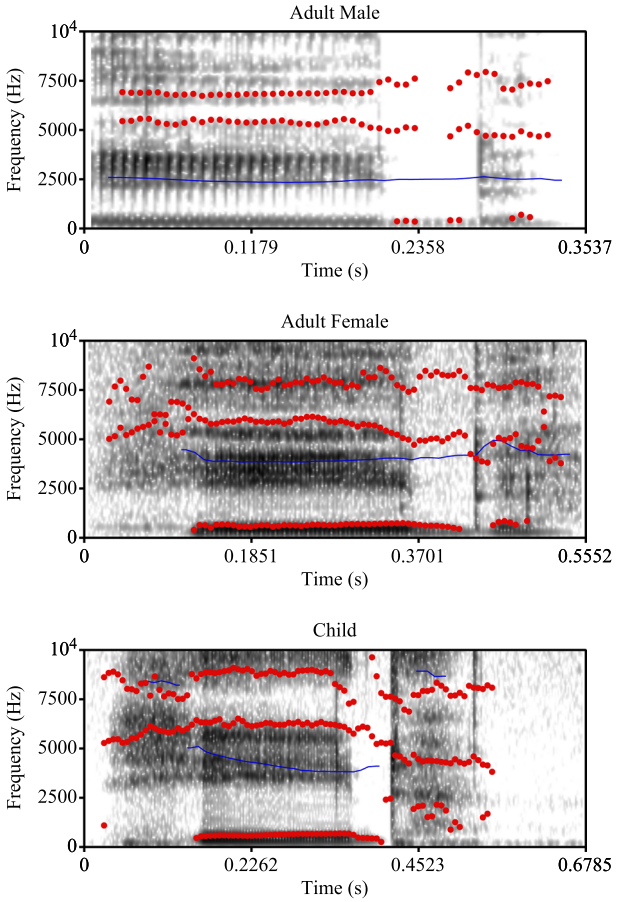

Fig. 4 Spectrograms corresponding to speech data from adult male (top panel), adult female (middle panel), and child (bottom panel) speaking the word *HEED*. The red speckles are the contours denoting the variation in formant frequencies, while the blue line denotes the pitch frequency variations (Color figure online)

necessary impetus toward the exploration for alternative front end acoustic features, namely the GTF-CC and the IGTF-CC.

3.2.1 Frame-Level Concatenation of MFCC and GTF-CC Feature Vectors

The Gamma-tone filter-banks are well known to better model the human auditory system [6, 15]. The Gamma-tone filter has a more smooth form and are placed in equal distance in frequency, in stark contrast with the Mel-filter-banks, as shown in the middle panel of Fig. 2. Moreover, the amount of overlap of a Mel-filter-bank is fixed, so if the number of filters increases, the bandwidth of each triangular filter will decrease. On the other hand, the bandwidth of a Gamma-tone filter is determined by its center frequency. So, if the number of filters increases, the overlap also increases. Theoretical and experimental results in [4] demonstrate that the filter bandwidth is one of the vital factors affecting speaker recognition performance in noise. Further the authors in [6], with the help of spectrograms, have analyzed the performance of ASV system subjected to a noisy speech utterance. The MFCC spectrogram is found to show robustness only at low frequencies. The GTF-CC spectrogram on the other hand showed robustness in both low and high frequencies, suggesting that GTF-CC features can play an influential role while dealing with child speakers. This has paved the idea of concatenating the MFCC and GTF-CC feature vectors to analyze the performance of short utterance-based children's ASV system.

The block diagram outlining the extraction process of the concatenated MFCC and GTF-CC features is shown in Fig. 3. The GTF-CC features are extracted in just the same way as the MFCC features discussed earlier, the only replacement being the Gamma-tone filter-bank in place of Mel-filter-bank. Both the MFCC features and the GTF-CC features are extracted using the Kaldi toolkit. Given the speech signal, first, we extract the MFCC and GTF-CC features. Next, for each of the short-time frames, the corresponding MFCC and GTF-CC features are appended. The resulting feature vectors (concatenated MFCC+GTF-CC feature vectors at the frame level) are then used as the input to the x -vector extraction process instead of the MFCC features. The experimental evaluations in the later portion of this paper demonstrate that an ASV system trained after concatenating GTF-CC features with MFCC features performs better than the one trained on MFCC features alone. However, it is worth highlighting that due to the inherent nature of filter-banks used in the feature concatenation of MFCC and GTF-CC, the spectral information in the higher-frequency range of the children's speech will be down-sampled. The quest for the preservation of higher-frequency contents in children's speech led us toward the exploration of another front-end acoustic feature, namely the IGTF-CC.

3.2.2 Frame-Level Concatenation of MFCC and IGTF-CC Feature Vectors

As already mentioned earlier, a significant amount of germane spectral information is present in the higher-frequency region in case of children. The spectrogram corresponding to speech data from children, in the bottom panel of Fig. 4, shows significant power even in the 4–8 kHz. Further to that, earlier literary works suggest that the formant frequencies are up-scaled in the case of child speakers [7, 12], which is also

quite prominently visible in the spectrogram of children speech in the bottom panel of Fig. 4. The spectrogram corresponding to speech data from adult male (top panel) and adult female (middle panel) is also plotted for comparison in Fig. 4. Mel-scale warping is inspired from the findings of psychoacoustics; it is based on the premise that human perception of pitch is linear up to 1000 Hz and then becomes nonlinear for higher frequencies (somewhat logarithmic) [3]. The Mel-filter-bank provides better resolution to speech signals in the low-frequency range, while its frequency resolution deteriorates in the high-frequency range, as illustrated by the nature of its filter-bank in the top panel of Fig. 2. When dealing with speech from children, the down-sampling of spectral information in the high-frequency range is a pitfall [7, 24]. Thus, preservation of spectral information in the higher-frequency range as well as the pursuit for filter-banks which best describes the human auditory system becomes our top-notch priority and persuades us to look for solutions beyond the traditional Mel-based filter-bank and Gamma-tone filter-bank for our high-pitched speakers. Motivated by this cognizance, the role of Inverse Gamma-tone filter-bank is delved into in this paper for the development of a robust children's ASV system.

The Inverse Gamma-tone filter-bank is realized simply by flipping around the Gamma-tone filters about the middle point of the frequency axis, as depicted in the bottom panel of Fig. 2. The front-end acoustic features achieved by replacing the Mel-filter-bank with Inverse Gamma-tone filter-bank are referred to as IGTF-CC features. This configuration of the filter-bank results in a better resolution of the spectral information in the high-frequency region, and thus, the Inverse Gamma-tone filter-bank is supposed to capture the acoustic information missed by the MFCC features. It is worth highlighting here that the Inverse Gamma-tone filter-bank is just a variant of Gamma-tone filter-bank, implying that it's filter-bank has the same smooth structure and whose bandwidth is decided by its center frequency just the same. So, if the number of filters increases, the overlap also increases. The Inverse Gamma-tone filter-bank though results in poor resolution to the lower-frequency components. Therefore, we have conceived the idea of concatenating the MFCC and IGTF-CC feature vectors in order to effectively preserve both the low- and high-frequency components. The block diagram outlining the extraction process of the concatenated MFCC and IGTF-CC features is also represented in Fig. 3. The IGTF-CC features are extracted in just the same way as the MFCC features discussed earlier, the only replacement being the Inverse Gamma-tone filter-bank in place of Mel-filter-bank. Both the MFCC features and the IGTF-CC features are extracted using the Kaldi toolkit. Given the fully augmented speech signal (employing the proposed out-of-domain augmentation technique), we firstly extract the MFCC and IGTF-CC features. Next, for each of the short-time frames, the corresponding MFCC and IGTF-CC features are appended. The resulting feature vectors (concatenated MFCC+IGTF-CC feature vectors at the frame level) are then used as the input to the x-vector extraction process instead of the MFCC features. The experimental evaluations in the later portion of this paper demonstrate that an ASV system trained after concatenating IGTF-CC features with MFCC features outperforms not only the ASV system trained on MFCC features alone, but also the one trained on the frame level fusion of MFCC and GTF-CC feature vectors.

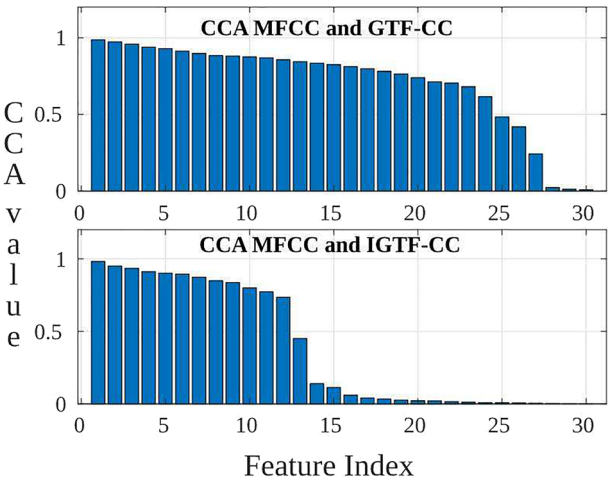

Fig. 5 Canonical correlation analysis of various feature concatenation explored in the paper

3.3 Canonical Correlation Analysis (CCA)

In order to substantiate the effect of feature concatenation, the canonical correlation analysis (CCA) was carried out. We have computed the canonical correlation among MFCC, GTF-CC, and IGTF-CC features as shown in Fig. 5. The CCA plot in the top panel of Fig. 5 shows how closely the MFCC and GTF-CC feature vectors are correlated for majority of the coefficients barring the last few coefficients. This explains for the inability of the concatenated MFCC and GTF-CC features to be able to capture the diverse range of acoustic attributes in children's speech. The CCA plot in the bottom panel of Fig. 5 shows that the MFCC and IGTF-CC feature vectors are highly uncorrelated or less correlated for most of the coefficients barring the starting few coefficients. Therefore, the frame-level concatenation of MFCC and IGTF-CC features leads to a wider range of acoustic attributes being captured. The inherently different configuration of filter-banks employed in the extraction of MFCC and IGTF-CC features is the main force behind this development. Thus, the CCA plot of IGTF-CC and MFCC reinforces the complementary characteristic of IGTF-CC with respect to MFCC which helps the duo in better capturing the acoustic information in children's speech.

4 Experimental Evaluations

In this section, the relative effectiveness of MFCC, concatenated MFCC and GTF-CC, concatenated MFCC and IGTF-CC features is explored and the experimentally verified results are presented.

4.1 The Speech Corpora

Four different speech corpora were employed for the development and evaluation of speaker verification system for children. These English-based speech corpus includes CSLU kids corpus [29], CMU kids corpus [5], PF-STAR kids corpus [1], and WSJ-CAM0 corpus [23].The details of each of these data sets are succinctly summarized as under, and a figurative tabulation of it is shown in Table 1:

1. *CSLU kids corpus*: This data set consists of spontaneous and prompted speech comprising of 100 h of data having 73, 100 utterances from 1100 children. The speech contribution is from children hailing from kindergarten to grade 10. Their speech data are sampled at a sampling rate of 16 kHz. This speech corpus is used as the training data for the ASV system in this work.
2. *CMU kids corpus*: This data set comprises of 9.1 h of data having 5180 utterances from 76 children. The child speakers are in the age group of 6–11 years. The sampling rate of this speech corpus is also 16 kHz, and it serves as our test set. A total of 423, 388 genuine trails and 26, 403, 832 impostor trails are present in this data set. The average duration of the data in this corpus is 6 s. Therefore, evaluation on this set represents the short utterance case.
3. *PF-STAR kids corpus*: It is an 8.3 h of data with utterances from 121 children speakers. The age group varies from 4 to 14 years. To maintain the uniformity of sampling rate with rest of our corpus, this data set has been down-sampled to 16 kHz from 44.1 kHz rate. This is our test data set for long utterances. It comprises 6, 664 genuine trails and 995, 420 impostor trails, respectively. The average duration of the data in this corpus is 30 s. Therefore, the evaluation on this set is for contrast in order to demonstrate the severity of the problem when short utterances are used.
4. *WSJCAM0 corpus*: It is an adults' speech corpus used for the out-of-domain data augmentation. This speech corpus comprises of 15.5 h of unperturbed data having 7852 utterances and 132, 778 words from 92 adult speakers (male and female). The sampling rate is 16 kHz.

4.2 Experimental Setup

The entire setup of the ASV system was developed and examined using the Kaldi toolkit [19]. In the process to extract the three kinds of aforementioned front-end features, speech data were first high-pass filtered having pre-emphasis factor of 0.97. It is a well-known fact that speech data are non-stationary in nature, so each of the speech utterances was first analyzed into short-time frames using overlapping Hamming windows. The duration of these overlapping Hamming windows was chosen to be 25 ms with a frame shift of 10 ms. For each of the three front-end features, a 30-channel filter-bank was employed to extract 30-dimensional base features. For MFCC features, a 30-channel Mel-filter-bank was engaged for warping the power spectra to Mel-scale, before computing the 30-dimensional MFCC features. For extracting the GTF-CC features, a 30-channel Gamma-tone filter-bank was engaged for warping the power spectra, before computing the 30-dimensional GTF-CC features. Finally, for the

Table 1 Details of different data sets used in this work for training and testing phase of the children's ASV system

Data used training	Duration of data(in hrs)	Number of speakers	Age range (in years)	Data used for testing	Duration of data(in hrs)	Number of speakers	Age range (in years)
CHILD(CSLU kids corpus)	100	1100	4–15	**CHILD(CMU kids corpus)**	**9.1**	**76**	6–11
ADULT(WSJCAM0 corpus)	15.5	92	≥18				
ADULT-VC	15.5	92	≥18				
ADULT-PM	31.5	184	≥18				
ADULT-FM-TSM	16	92	≥18				
CHILD + ADULT + ADULT-FM-TSM + ADULT-PM + ADULT-VC	**178.5**	**1560**	NA				

Bold values indicate the complete dataset used for training and testing phase of the children's ASV system

computation of the IGTF-CC features, a 30-channel Inverse Gamma-tone filter-bank was superimposed over the power spectrum, before computing the 30-dimensional IGTF-CC features.

Description of out-of-domain data augmentation: The out-of-domain training set used for developing the children's ASV system was derived from an adult's speech corpus called as WSJCAM0 corpus. This training data set consists of original adult speech data derived from both male and female speakers and is referred to as ADULT. Three newer versions of speech data are synthetically generated from this speech corpora and are enlisted as follows:

i. ADULT-VC: This data set was generated by applying voice conversion to the adult data through a cycle-consistent generative adversarial network (C-GAN). The GAN underwent training using a 10-min speech data set encompassing both adult (source) and child speakers (target). The number of epochs utilized in training the C-GAN parameters was set at 5000;
ii. ADULT-PM: This data set was generated by increasing the duration of the speech data of ADULT by a factor of 1.4 while the pitch of ADULT was enhanced by a factor of 1.35. To perform time-scale modification, the technique of audio stretching was applied, leveraging the methodology of fuzzy classification of spectral bins (FCSB) [2];
iii. ADULT-FM-TSM: This data set was generated by up-scaling the formant frequencies of ADULT speech by a factor of 0.08. At the same time, the speaking rate of adults' speech data was decreased by a factor of 1.4 through time-scale modification.

After performing the aforementioned data modification techniques namely, voice conversion (VC), prosody modification (PM), formant and time-scale modification (FM-TSM), a total of 63 h of synthetic data was available for training purpose with acoustic attributes similar to those of children's speech.

For the extraction of highly discriminative speaker representations, a deep neural network was utilized. These fixed-dimensional speaker embeddings called as *x*-vectors were extracted from a time-delay neural network (TDNN) architecture [16, 30, 31]. This architecture consists of 7 hidden layers and undergoes training for 6 epochs. The TDNN architecture is structured into three integral components: the frame-level, statistics-level, and segment-level components. Within the frame-level component, spanning layers 1–5, input features sequentially traverse these layers, effectively capturing temporal information and enhancing the temporal context of the frames under consideration. The statistics-level component serves the purpose of converting variable-length speech inputs into a singular, fixed-dimensional vector. This component consists of a single layer, called the statistics pooling, which amalgamates the output vectors from the TDNN's frame-level and computes their mean and standard deviation. Concurrently, the segment-level component is responsible for attributing speaker identities to the segment-level vector. The mean and standard deviation, post-concatenation, are transmitted to two additional hidden layers, subsequently leading to a softmax output layer. Layer 6 operates as the speaker embedding, which transforms the information from the preceding layer into a low-dimensional representation. This intricate arrangement of components and layers underscores the comprehensive

design of the TDNN architecture and its efficacy in processing speech inputs at varying levels of abstraction. The training of network parameters was conducted utilizing the stochastic natural gradient descent algorithm [20, 31]. Finally, each of the speech utterances was represented as a 512-dimensional x-vector. The scoring process was executed through the utilization of x-vectors in conjunction with the trained PLDA model. When provided with two per-utterance embeddings, denoted as e_i and e_j, the PLDA computes a log-likelihood ratio (LLR) to quantify the likelihood associated with the pair of embeddings. The LLR is calculated in the following manner:

$$\mathrm{LLR}(e_i, e_j) = \log \left[\frac{P\left(\frac{e_i, e_j}{H_1}\right)}{P\left(\frac{e_i, e_j}{H_0}\right)} \right] \tag{1}$$

where H_1 represents the hypothesis related to the same speaker, while H_0 pertains to the hypothesis associated with different speakers. The PLDA model calculates a log-likelihood ratio for each speaker pair, representing the level of similarity between the individuals. In instances where the pair shares the same label, a high score is anticipated, signifying identical speakers (a genuine claim). Conversely, when the pair bears different labels, a low score is expected, indicating different speakers (an imposter). The metrics used for performance measure were equal error rate (EER) and minimum decision cost function (minDCF).

4.3 Experimental Results

This study was carried out to monitor how the performance of an ASV system, trained on a mix of a large amount of children's speech data and an adequate amount of original as well as modified adult's speech corpus, is affected when subjected to short utterances of children's speech. The EER and minDCF values for the employed ASV system are given in Table 2. When subjected to short utterances, a relative improvement of 33.6% with respect to the baseline system trained solely on child data set is achieved when the proposed data augmentation techniques are applied. This shows that the proposed data augmentation technique is very effective. The EER and minDCF values, when the employed ASV system is tested with long utterances of children's speech test set, are also enlisted for comparison. As can be seen from Table 2, the EER of the baseline system climbs from 6.38% (for long test utterances) to 21.95% (for short test utterances). Further, when the proposed out-of-domain data augmentation has been employed, the EER of ASV system climbs from 3.824% (for long test utterances) to 14.58% (for short test utterances). This shows the magnanimity of the challenge posed by short test utterances on the ASV performance. At the same time, it is imperative to realize and appreciate that the proposed data augmentation approach takes the edge off the detrimental effect of short utterance speech test set.

Next, the effectiveness of the frame-level concatenation of the front-end acoustic features in the light of the employed short utterance-based ASV system was examined. The EER and minDCF values obtained when MFCC and GTF-CC features are concatenated, as well as for MFCC and IGTF-CC feature fusion given in Table 3. In this

Table 2 EER and minDCF values for the short and long utterances of children's speech test set demonstrating the effectiveness of out-of-domain data augmentation techniques

Data used for training	Short utterances		Long utterances	
	EER(%)	minDCF	EER(%)	minDCF
CHILD (Baseline)	21.95	0.9975	6.38	0.7228
CHILD + ADULT + ADULT-FM-TSM + ADULT-PM + ADULT-VC (Proposed)	**14.58**	**0.9233**	**3.82**	**0.4062**

Bold values indicate the lower values of EER obtained for the proposed data augmentation technique and proposed feature concatenation approach respectively
The out-of-domain data augmentation scheme includes adult voice conversion (ADULT-VC), adult formant and time-scale modification (ADULT-FM-TSM), adult prosody modification (ADULT-PM)

Table 3 EER and minDCF values for the short utterance-based ASV system trained on the data set obtained using the proposed out-of-domain data augmentation technique demonstrating the effectiveness of feature concatenation

Acoustic features	Evaluation metric	
	EER (%)	minDCF
MFCC	14.58	0.9233
MFCC + GTF-CC	13.96	0.9617
MFCC + IGTF-CC	**13.50**	**0.9041**

Bold values indicate the lower values of EER obtained for the proposed data augmentation technique and proposed feature concatenation approach respectively

case, the proposed data augmentation technique has been employed prior to training the ASV system. The EER and minDCF values obtained when MFCC features are used alone are also enlisted for comparison. As evident, an absolute reduction of 0.62% in EER is achieved by concatenation of MFCC with GTF-CC features, while the concatenation of MFCC and IGTF-CC features yields an absolute reduction in EER of 1.08%. The detection error trade-off (DET) plot summarizing these results is shown in Fig. 6. In this plot, baseline refers to the ASV trained exclusively on children's speech using MFCC features.

To access the effectiveness of the proposed approach in a more comprehensive manner, an age-wise analysis as well as gender-wise analysis of children's speech was performed. For evaluating the effect of age variation, the evaluation metric results are reported for the entire test set, as well as after doing the age-wise break-up of the test set in two subgroups. The EER and minDCF values for this experimental study are given in Table 4. In this case as well, the proposed out-of-domain data augmentation has been employed before training the *x*-vector extractor. The first subgroup comprised speech utterances from speakers belonging to the age-group 6–7 years while the second subgroup comprised of speech utterances of speakers hailing from the age-group 8–9 years. As evident from the enlisted results, a significant degradation (reflected in the poor values of EER) is noted for children in the lower age-group (6–7 years) compared to the children in the higher age-group (8–9 years) or against the children in the full test set. This degradation is due to higher formant frequency and pitch frequency of children's speech due to their inherent shorter vocal tract length. As the children grow, the formant frequencies decrease as well as the speaking rate tends to stabilize.

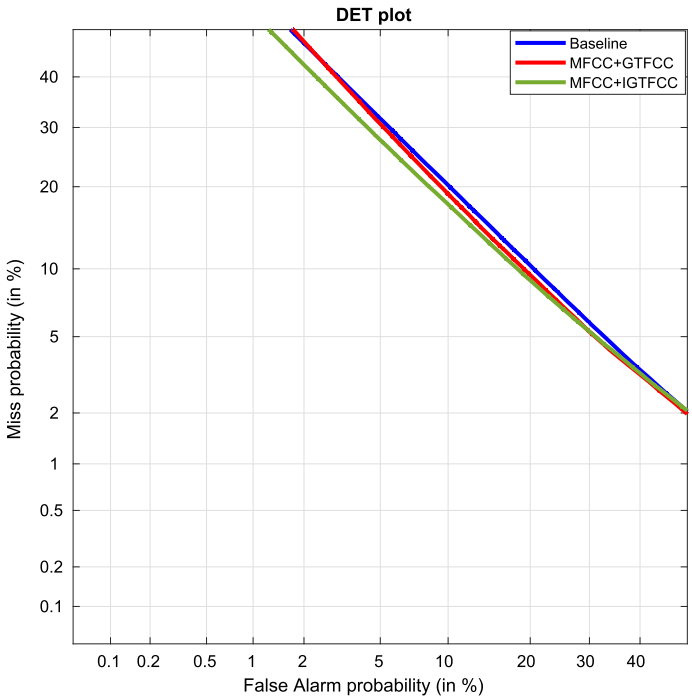

Fig. 6 Detection error trade-off plot demonstrating the effectiveness of proposed feature concatenation

Table 4 Age group-wise break up of EER and minimum DCF values highlighting the significance of feature concatenation approaches

Features	Age group(in years)	EER (%)	minDCF
MFCC	Full test set	14.58	0.9233
	6–7	17.39	0.9657
	8–9	14.04	0.9203
MFCC+GTF-CC	Full test set	13.96	0.9617
	6–7	16.21	0.9915
	8–9	13.22	0.9552
MFCC+IGTF-CC	Full test set	**13.50**	0.9041
	6–7	**15.42**	0.9547
	8–9	**12.88**	0.9035

This study was performed on *x*-vector-based ASV system trained on a mix of children's speech and adults' speech along with the modified versions of adults' speech

Further, it is noteworthy that the ASV system trained solely on the MFCC features performs poorly in terms of evaluation metrics as it down-samples the higher-frequency contents of children's speech. The children's ASV system trained on the concatenated acoustic features yields better results, and this improvement is more profound in the lower age-group as the concatenation of GTF-CC/IGTF-CC features with

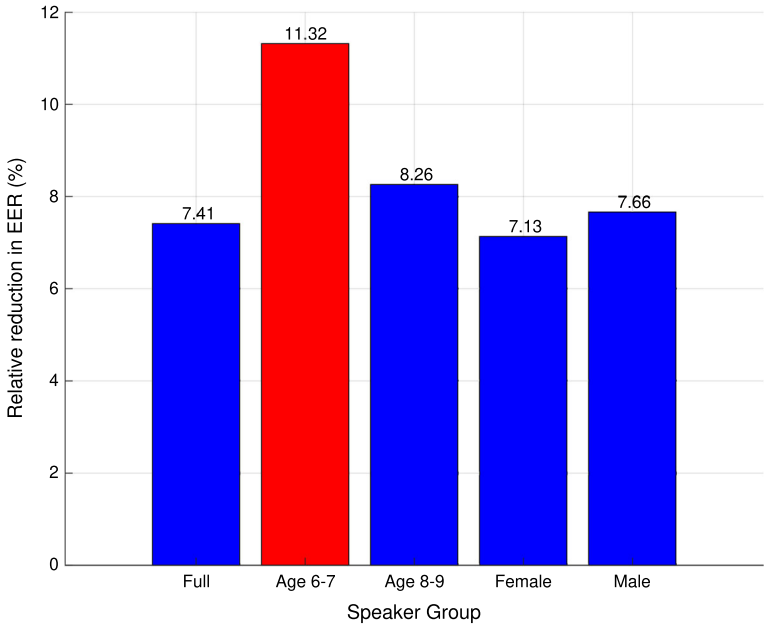

Fig. 7 Bar graph representation of the relative reduction in EER(%) for various speaker groups(in terms of age and gender) corresponding to the ASV system trained on the concatenated MFCC and IGTF-CC features as compared to an ASV system trained on the MFCC features alone. The bar depicted in red shows the greatest relative improvement in EER

the MFCC features takes into account the spectral information in the lower- as well as higher-frequency regions. The EER for the full test set shows a relative reduction of 7.41% when MFCC and IGTF-CC features are concatenated, pictorially depicted by the first bar in Fig. 7. The corresponding relative reduction in error of the same concatenated features for the age group 6–7 years is **11.32%**, pictorially depicted by the second bar in Fig. 7. For the age group between 8 and 9 years, the relative reduction in EER is 8.26%, pictorially depicted by the third bar in Fig. 7.

Finally, the effect on the performance of the employed ASV system was evaluated due to gender-wise grouping of the children's speech test set. The EER and minDCF values for this experimental study are given in Table 5. As evident from the enlisted results, a significant degradation (reflected in the poor values of EER) is noted for the female children as compared to the male children or when compared with the children in the full test set. This degradation is due to higher formant frequencies of female children's speech compared to their male counterparts. Further, as evident from the table, the EER considerably reduces when either GTF-CC features or IGTF-CC features are concatenated with MFCC features. The EER for the full test set shows a relative reduction of 7.41% when MFCC and IGTF-CC features are concatenated, pictorially depicted by the first bar in Fig. 7. The corresponding relative reduction in error of the same concatenated features for the female child is 7.13%, pictorially depicted by the fourth bar in Fig. 7. For the male child, the relative reduction in EER is 7.66% as compared to the baseline, pictorially depicted by the fifth bar in Fig. 7.

Table 5 Gender-wise breakup of EER and minimum DCF values highlighting the significance of feature concatenation approaches

Features	Child Gender	EER (%)	minDCF
MFCC	Full test set	14.58	0.9233
	Female	17.68	0.9585
	Male	10.05	0.8534
MFCC+GTF-CC	Full test set	13.96	0.9617
	Female	17.12	0.9929
	Male	**9.18**	0.8548
MFCC+IGTF-CC	Full test set	**13.50**	0.9041
	Female	**16.42**	0.9466
	Male	**9.28**	0.8055

This study was performed on *x*-vector-based ASV system trained on a mix of children's speech and adults' speech along with the modified versions of adults' speech

5 Conclusion and Future Research Direction

The work in this paper sets forth our endeavor toward the development of a robust children's ASV system using short utterances under low-resource conditions. To address the inevitable problem of speech data paucity, an out-of-domain data augmentation technique is proposed to synthetically generate more data for training. Out-of-domain data augmentation approach helps in widening the diversity of the captured acoustic attributes, by introducing missing desirable characteristics while keeping the acoustic mismatch in check. Together with data augmentation, the effectiveness of frame-level concatenation of MFCC with the GTF-CC/ IGTF-CC is also analyzed in this paper. In GTF-CC or its variant, the IGTF-CC features are well known to better model the human auditory system and are more resilient to additive noise compared to the traditional MFCC features. Additionally, the complementary nature of filter-bank in the IGTF-CC with respect to MFCC helps in preserving spectral information in the higher-frequency range. Thus, MFCC features in tandem with the IGTF-CC features help not only in modeling the human auditory model in a more competent manner, but also in preserving the spectral information in low- as well as high- frequency range. Furthermore, age- and gender-wise analyses were carried out to study the combined effect of data augmentation and feature concatenation on the ASV system performance. Children in the lower age bracket exhibit more pronounced inter-speaker variability, resulting in a degraded performance in terms of EER and minDCF compared to the children in the higher age bracket. At the same time, the employed ASV system incorporating both the proposed data augmentation technique and feature concatenation is found to be more impactful for children in the lower age group, resulting in a significant reduction in EER and minDCF compared to the baseline.

As a future extension of this work, in addition to the out-of-domain data derived from adults' speech, we would like to explore the effectuality of in-domain data augmentation techniques for the purpose of increasing the amount and diversity of the captured acoustic attributes of children's speech for training. In-domain data augmentation refers to increasing the amount of children's speech available for training

by synthetically generating more data from children's speech itself. In this regard, we would like to implement speed perturbation and pitch perturbation of the original children's speech. In addition, we would also like to explore and incorporate the vocal tract length perturbation (VTLP) technique. VTLP approach explicitly models and compensates for the ill-effects of variations in vocal tract length by introducing diversity into the complete children speech data set by creating numerous sets of data with varying linear warping factors. The out-of-domain data augmentation techniques in tandem with the in-domain data augmentation techniques are anticipated to reduce the EER and minDCF values, which will eventually help in the realization of a more robust and dependable children's ASV system.

Funding Funding information is not applicable / no funding was received.

Data Availability The data sets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical Approval The work presented in the uploaded manuscript is an original one, and the manuscript is not currently under consideration for publication elsewhere.

Consent for Publication It is hereby confirmed that the manuscript has been read and approved for submission by all the named authors. It is therefore requested to consider the submitted manuscript for publication in the esteemed journal.

References

1. 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. Proceedings INTERSPEECH, pp. 2761–2764 (2005)
2. E.P. Damskägg, V. Välimäki, Audio time stretching using fuzzy classification of spectral bins. Appl. Sci. **7**(12), 1293 (2017)
3. 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). https://doi.org/10.1109/TASSP.1980.1163420
4. D. Dimitriadis, P. Maragos, A. Potamianos, On the effects of filterbank design and energy computation on robust speech recognition. IEEE Trans. Audio Speech Lang. Process. **19**(6), 1504–1516 (2010)
5. M. Eskenazi, J. Mostow, D. Graff, The CMU Kids Corpus LDC97S63. https://catalog.ldc.upenn.edu/LDC97S63 (1997)
6. M. Fedila, M. Bengherabi, A. Amrouche, Gammatone filterbank and symbiotic combination of amplitude and phase-based spectra for robust speaker verification under noisy conditions and compression artifacts. Multimed. Tools Appl. **77**(13), 16721–16739 (2018)
7. M. Gerosa, D. Giuliani, S. Narayanan, A. Potamianos, A review of ASR technologies for children's speech. Proceeding Workshop on Child, Computer and Interaction, pp. 7:1–7:8 (2009)
8. B. Gold, N. Morgan, D. Ellis, *Speech and audio signal processing: processing and perception of speech and music* (Wiley, 2011)
9. B. Gold, N. Morgan, D. Ellis, D. O'Shaughnessy, Speech and audio signal processing: processing and perception of speech and music, second edition. J. Acoust. Soc. Am. **132**, 1861–2 (2012). https://doi.org/10.1121/1.4742973

10. T. Kaneko, H. Kameoka, Parallel-data-free voice conversion using cycle-consistent adversarial networks. arXiv preprint arXiv:1711.11293 (2017)
11. H.K. Kathania, S.R. Kadiri, P. Alku, M. Kurimo, Study of formant modification for children asr. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (IEEE, 2020), pp. 7429–7433
12. H.K. Kathania, S. Shahnawazuddin, W. Ahmad, N. Adiga, Role of linear, mel and inverse-mel filterbanks in automatic recognition of speech from high-pitched speakers. Circuits Syst. Signal Process. **38**(10), 4667–4682 (2019)
13. V. Kumar, A. Kumar, S. Shahnawazuddin, Creating robust children's ASR system in zero-resource condition through out-of-domain data augmentation. Circuits Syst. Signal Process. **41**(4), 2205–2220 (2022)
14. S. Lee, A. Potamianos, S.S. Narayanan, Acoustics of children's speech: developmental changes of temporal and spectral parameters. J. Acoust. Soc. Am. **105**(3), 1455–1468 (1999)
15. R.D. Patterson, I. Nimmo-Smith, J. Holdsworth, P. Rice, An efficient auditory filterbank based on the gammatone function. A meeting of the IOC Speech Group on Auditory Modelling at RSRE, vol. 2 (1987)
16. V. Peddinti, D. Povey, S. Khudanpur, A time delay neural network architecture for efficient modeling of long temporal contexts. Proceedings INTERSPEECH (2015)
17. A. Poddar, M. Sahidullah, G. Saha, Speaker verification with short utterances: a review of challenges, trends and opportunities. IET Biom. **7**(2), 91–101 (2018)
18. A. Poddar, M. Sahidullah, G. Saha, Quality measures for speaker verification with short utterances. Digit. Signal Process. **88**, 66–79 (2019). https://doi.org/10.1016/j.dsp.2019.01.023
19. 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. Proceedings ASRU (2011)
20. D. Povey, X. Zhang, S. Khudanpur, Parallel training of deep neural networks with natural gradient and parameter averaging. Proceedings ICLR (2015)
21. S.R.M. Prasanna, D. Govind, K.S. Rao, B. Yegnanarayana, Fast prosody modification using instants of significant excitation. Proceedings International Conference on Speech Prosody (2010)
22. P. Rajan, T. Kinnunen, C. Hanilci, J. Pohjalainen, P. Alku, Using group delay functions from all-pole models for speaker recognition. INTERSPEECH, pp. 2489–2493 (2013)
23. T. Robinson, J. Fransen, D. Pye, J. Foote, S. Renals, WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition. Proceedings ICASSP **1**, 81–84 (1995)
24. M. Russell, S. D'Arcy, Challenges for computer recognition of children's speech. Proceedings Speech and Language Technologies in Education (SLaTE) (2007)
25. S. Safavi, M. Russell, P. Jančovič, Automatic speaker, age-group and gender identification from children's speech. Comput. Speech Lang. **50**, 141–156 (2018)
26. S. Shahnawazuddin, N. Adiga, H.K. Kathania, B.T. Sai, Creating speaker independent ASR system through prosody modification based data augmentation. Pattern Recogn. Lett. **131**, 213–218 (2020). https://doi.org/10.1016/j.patrec.2019.12.019
27. S. Shahnawazuddin, N. Adiga, B.T. Sai, W. Ahmad, H.K. Kathania, Developing speaker independent ASR system using limited data through prosody modification based on fuzzy classification of spectral bins. Digit. Signal Process. **93**, 34–42 (2019)
28. S. Shahnawazuddin, W. Ahmad, N. Adiga, A. Kumar, In-domain and out-of-domain data augmentation to improve children's speaker verification system in limited data scenario. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7554–7558 (2020)
29. K. Shobaki, J.P. Hosom, R. Cole, Cslu: Kids' speech version 1.1. Linguistic Data Consortium (2007)
30. D. Snyder, D. Garcia-Romero, D. Povey, S. Khudanpur, Deep neural network embeddings for text-independent speaker verification. Proceedings INTERSPEECH, pp. 999–1003 (2017)
31. D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, X-Vectors: Robust DNN Embeddings for Speaker Recognition. Proceedings ICASSP, pp. 5329–5333 (2018)
32. G. Yeung, A. Alwan, On the difficulties of automatic speech recognition for kindergarten-aged children. Interspeech 2018 (2018)