Circuits, Systems, and Signal Processing (2024) 43:1715–1740
https://doi.org/10.1007/s00034-023-02535-8

Short-Utterance-Based Children's Speaker Verification in Low-Resource Conditions

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

Received: 22 December 2022 / Revised: 7 October 2023 / Accepted: 8 October 2023 /
Published online: 5 November 2023

Abstract

The task of developing an automatic speaker verification (ASV) system for children is extremely challenging due to unavailability of sufficiently large and free speech corpora from child speakers . On the other hand, hundreds of hours of speech data from adult speakers are freely available. Therefore, majority of the works on speaker verification reported in the literature deal predominantly with adults' speech, while only a few works dealing with children's speech have been published. The challenges in developing a robust ASV system for child speakers are further exacerbated when we use short utterances which is largely unexplored in the case of children's speech . Therefore, in this paper, we have focused on children's speaker verification using short utterances. To deal with data scarcity, several out-of-domain data augmentation techniques have been utilized. Since the out-of-domain data used in this study is from adult speakers which is acoustically very different from children's speech, we have resorted to 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 relevant to children's speech. A staggering relative improvement of 33.57% in equal error rate with respect to the baseline system trained solely on child dataset speaks volume of the effectiveness of the proposed data augmentation technique in this paper. Further to that, we have also proposed frame-level concatenation of Mel-frequency cepstral coefficients (MFCC) with frequency-domain linear prediction coefficients, in order to simultaneously model the spectral as well

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

Ankita
ankita.ph21.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

as temporal envelopes. The proposed idea of frame-level concatenation is expected to further enhance the discrimination among the speakers. This novel approach, when combined with data augmentation, helps in further improving the performance of the speaker verification system. The experimental results support our claims, wherein we have achieved an overall relative reduction of 38.04% for equal error rate.

Keywords Automatic speaker verification · Out-of-domain data augmentation · Prosody modification · Formant modification · Feature concatenation · Frequency-domain linear prediction

1 Introduction

Automatic speaker recognition is the process of automatically recognizing a speaker from his/her voice samples. Speaker recognition is divided into two main activities viz. speaker identification and speaker verification [26]. In the case of speaker identification, the aim is to identify a speaker from among either a closed or an open set of speakers given the test speech sample. On the other hand, automatic speaker verification (ASV) addresses the poignant authentication issue of the claimed identity of a speaker. In this process, a speaker feeds in his/her data and claims the identity of a particular person. The deployed ASV system then digs into its stored database and matches the earlier learned template of the claimed identity with the input speech sample. The ASV system then pronounces the claim to be genuine if the test speech samples and the stored template match to a certain permissible degree; else the speaker is declared an impostor. An ASV system can be further categorized into 'text-dependent' and 'text-independent' depending upon whether the word or sentence level transcriptions of the speech inputs are used while developing the ASV system.

In the past quinquennial, social networking websites and e-learning tools have become the new normal among people of all age groups. These technological marvels come as a package, which is fraught with dangers of losing sensitive data and identity theft. Those keeping abreast with it should be wary of these lurking perils. Even though such a menacing issue can victimize anyone, children who are less aware of the magnanimity surrounding the perils of losing sensitive data and identity theft become more vulnerable targets. To address such an intimidating issue, a plethora of security measures are being deployed; automatic speaker verification system is one among those. The necessity of a robust ASV systems considering their pivotal role in providing security and protection, entertainment, games and education, surveillance [14] has thus taken a center stage in these testing times and is bound to grow by leaps and bounds in the years to come. The focus of the researchers around the globe are thus riveted on the development of an ASV system that can ascertain the speaker's identity with a low error rate. Dismally, the major chunk of the works reported in the literature hover around with the task of building an ASV system for adult population. The literary works reported on building an ASV system for children are regrettably minimal [20, 26, 33]. Motivated by these facts, the work presented in this paper focuses on developing an ASV system for child speakers.

1.1 Challenges

Building a state-of-the-art children ASV systems is plagued by several challenges and no wonder why literary works reported in this domain is minimal. The lack of sizable speech corpora which are freely available act as the first roadblock. Further, children speech corpora are available in only a handful of languages spoken across the globe [26]. 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 offering (low-resource condition), developing a children's ASV system, exploiting recently proposed techniques employing deep learning architectures is still very challenging. State-of-the-art ASV systems incorporate deep learning architectures that require estimating a large number of parameters. This, in turn, requires a large amount of domain-specific data. To circumvent with the low- and zero-resource conditions, a few earlier works on children's ASV have performed extensive studies on the impact of synthetically generating speech data that have acoustic attributes similar to that of children's speech and then pooling it into training. Out-of-domain data augmentation has been reported to be effective in the context of children's ASV task [25].

The performance of an ASV system for children is further dented when there is a reduction in the duration of the test speech utterances, commonly termed as short-utterance situations. Speech segments of duration 5–10 seconds are commonly termed as short-utterances in the literary domain. The unavailability of sufficiently longer duration of speech data can be tackled during training phase by some data augmentation techniques; it is not feasible to do the same during the testing phase [14]. The works reported on children's ASV hardly deal with such short-utterance scenario.

1.2 Proposed Approaches

Motivated by the facts discussed earlier, we have studied the role of out-of-domain data augmentation in the context of short-utterance-based children's ASV task. In this regard, we have explored the effect of synthetically generating speech data which is acoustically similar to that of children's speech from the available adult speech corpus prior to augmentation. The techniques used to address the dearth of domain-specific data explored in this paper includes (i) voice conversion (VC) of adults' speech data through a cycle-consistent generative adversarial network (C-GAN) [8], (ii) prosody modification (PM) [22, 24] of adults' speech, i.e., optimally changing the pitch and duration of the speech data from adult speakers, and (iii) up-scaling the formant frequencies (FM) [9, 12] 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. Therefore, the explored out-of-domain data augmentation techniques are observed to be very effective as demonstrated through the experimental studies presented in this paper.

In general, the Mel-frequency cepstral coefficients (MFCC) are the most widely used front-end acoustic features in the context of speaker verification task. However,

in this study, we have also explored the role of another well-known front-end speech parameterization technique namely frequency-domain linear prediction (FDLP) coefficients. The MFCC features capture the spectral envelope. On the other hand, the FDLP features capture the temporal envelope. In this paper, we have explored the efficacy of FDLP in isolation as well as in combination with MFCC features in the context of children's speaker verification task. Our experimental explorations show that simultaneous modeling of temporal as well as spectral envelopes (i.e., frame-level concatenation of MFCC and FDLP features) leads to better results compared to the case when either MFCC or FDLP features are used. Furthermore, we have also studied the effectiveness of employing three different filter-banks while extracting the FDLP features. The studied filter-banks are the Mel-, Bark-, and linear filter-banks. The presented experimental evaluations show that using linear filter-bank while extracting the FDLP features yields superior results. This is because of the fact that a significant amount of relevant spectral information is present in the higher-frequency components in the case of children's speech that get averaged out when either the Mel- or Bark-filter-banks are used. On the other hand, the use of linear filter-bank helps in effectively preserving the information present in the higher-frequency components.

As demonstrated later in this paper, the proposed frame-level concatenation of the two kinds of front-end acoustic features increases the class separation among the speakers. This, in turn, enhances the performance with respect to children. The experimental evaluations demonstrate that an ASV system trained after concatenating FDLP features with MFCC features outperforms the one trained on MFCC/FDLP alone. The paper also elucidates the age-group as well as gender wise analysis of system performance to figure out the consequences of data augmentation and feature concatenation. This proposed approach aids in considerably subsiding the equal error rate (EER) and detection cost function (DCF) as opposed to our baseline system trained exclusively on children's speech using MFCC features. The ASV system for children's speech incorporated in this work for experimental evaluations employs *x*-vector-based speaker representation along with probabilistic linear discriminant analysis (PLDA) based scoring.

The remainder of this paper is organized as follows: Section 2 presents a comprehensive literature review of some of recent papers underlining the importance of representing temporal structures using FDLP in ASV related tasks. Section 3 describes the proposed out-of-domain data augmentation techniques to deal with scarcity of domain-specific data. In Sect. 4, we have cast light upon the motivation of frame-level concatenation of FDLP and MFCC features in our proposed children ASV system. The experimental evaluations exhibiting the efficacy of our proposed technique are presented in Sect. 5. In Sect. 6, an overview of the discussed approaches are tabulated along with the advantages and limitations of each discussed approach. Section 7 talks about the future scope of this work. Eventually, the conclusion in Sect. 8 brings down the curtain on the paper.

2 Related Works

Spectrum-based analysis techniques are traditionally used in automatic speech and speaker recognition for acoustic modeling. Spectral structures such as formants do convey essential linguistic details but its only a partial representation of speech signals. On the other hand, temporal structure in the sub-10 ms transient segment contains crucial indications for both the perception of natural sounds and the comprehension of speech stop bursts. This underlines the importance of capturing the temporal envelop along with spectral envelop in the performance of speaker verification tasks. Earlier works have shown that modeling the temporal envelop helps in enhancing the performance of tasks such as recognizing reverberant speech [31], replay spoofing attack detection [32], spoken term detection [13] and speech synthesis attacks [21]. ASV systems are susceptible to a number of spoofing attacks, including replay, voice conversion, and speech synthesis. In [32], the authors' presented a study on temporal envelope features for the detection of replay spoofing attacks, which were extracted using the FDLP framework.

The study in [10] proposes to use FDLP coefficients for dialect classification motivated by its long temporal summarization during pole estimation. Support vector machine (SVM) and feed-forward neural network (FFNN) classifiers use the i-vectors and x-vectors derived from both baseline (MFCCs, linear prediction cepstral coefficients (LPCCs), perceptual LPCCs (PLPCCs), RASTA filtered PLPCCs (PLPCC-R)), and FDLP features to identify the dialects. It has been demonstrated in that study that FDLP coefficient features outperform baseline features like MFCCs and PLPCC. Additionally, it is demonstrated that the baseline features and the FDLP features include complementary information.

Language identification systems performance degrades on account of a mismatch between training and testing speech utterances, especially when dealing with short duration utterances. The idea that long-term trends are less impacted by this mismatch than short-term features is explored in [5]. It specifically suggests using characteristics based on the temporal envelopes of sub-bands. In that study, linear prediction in the frequency domain is used to get the temporal envelopes. The cepstral characteristics are then created from those envelopes. A bidirectional long short-term memory recurrent neural network is then employed in order to identify languages. Experimental analyses shows that in comparison with baseline features, the proposed features display significantly improved robustness under various noise and mismatch conditions. Particularly, across the test-set, the proposed features outshine cutting-edge bottleneck characteristics.

The speech activity detection (SAD) technique for speaker verification in noisy contexts is presented in [6]. The phoneme posteriors obtained from a multi-layer perceptron (MLP) are the foundation of the proposed SAD. In order to train the MLP, long temporal chunks of the speech stream were examined in key bands utilizing modulation spectral characteristics. FDLP was used to determine the temporal envelopes for each sub-band. A minimum mean square envelope estimate technique produced sub-band envelopes that were resilient. The trained MLP used the speech features as input to calculate phoneme posterior probabilities. To determine speech/non-speech decisions for SAD, all speech phoneme probabilities were combined into a single

speech class. The suggested SAD was used for a speaker verification task employing noise degraded versions of the NIST 2008 speaker recognition evaluation data, and it significantly reduced the relative equal error rate and hence enhanced the performance of the ASV task.

The work in [7] presents a strong text-dependent speaker identification system based on the unique FDLP feature. With the use of a 2-D regressive model, this feature was extracted from the all pole system. Performances of the proposed approach were evaluated in both quiet and noisy environments. Under noisy situations, the introduced technique performed better than all competing methods for all studied signal to noise ration (SNR) values, and its performance in clean conditions was equivalent. The suggested method also produced a consistent pattern across many noise types and offered very reliable performance. The findings of this study reveal that the suggested method may simulate the human voice production system quite precisely, which accounts for its robustness.

It is imperative to realize and appreciate that all the aforementioned works on FDLP-based ASV system are built for adult speakers and none of them addresses the challenges involved in building short-utterance-based children's ASV system. This motivated us to study the impact of employing FDLP features in the case of children's ASV task using short-utterances. To the best of our knowledge, the effectiveness of FDLP features in the context of children's ASV task has not been studied yet. The work presented in this paper studies the role of FDLP features not only in isolation but also in combination with MFCC features. The experimental evaluations presented in this paper suggest that simultaneous modeling of both spectral as well as temporal envelopes is more powerful than modeling either of the two in isolation. Hence, frame-level concatenation of MFCC and FDLP features is proposed in this study.

3 Data Augmentation

As mentioned earlier, state-of-the-art ASV systems employ *x*-vectors-based speaker representation. For extracting the *x*-vectors, a time-delay neural network (TDNN) comprising of a large number of hidden layers and hidden nodes per layer is trained. When the training data is amply large, the fixed length vector representation derived from the speech data (aka *x*-vectors) are reported to be immensely effectual. As already discussed, one of the hindrances in the development of a reliable ASV systems for children is a dearth of domain-specific data. Therefore, training an *x*-vector-based ASV system on a limited amount of children's speech results in a lackluster performance. Out-of-domain data augmentation can help alleviate this problem. Motivated by this, we have resorted to synthetically generating speech data which is acoustically similar to that of children's speech using the available adults' speech corpus. The synthetically generated data were then pooled into training in order to circumvent the detrimental effect of data scarcity.

Several ways of data augmentation have been explored in this work and are briefly explained in the following:

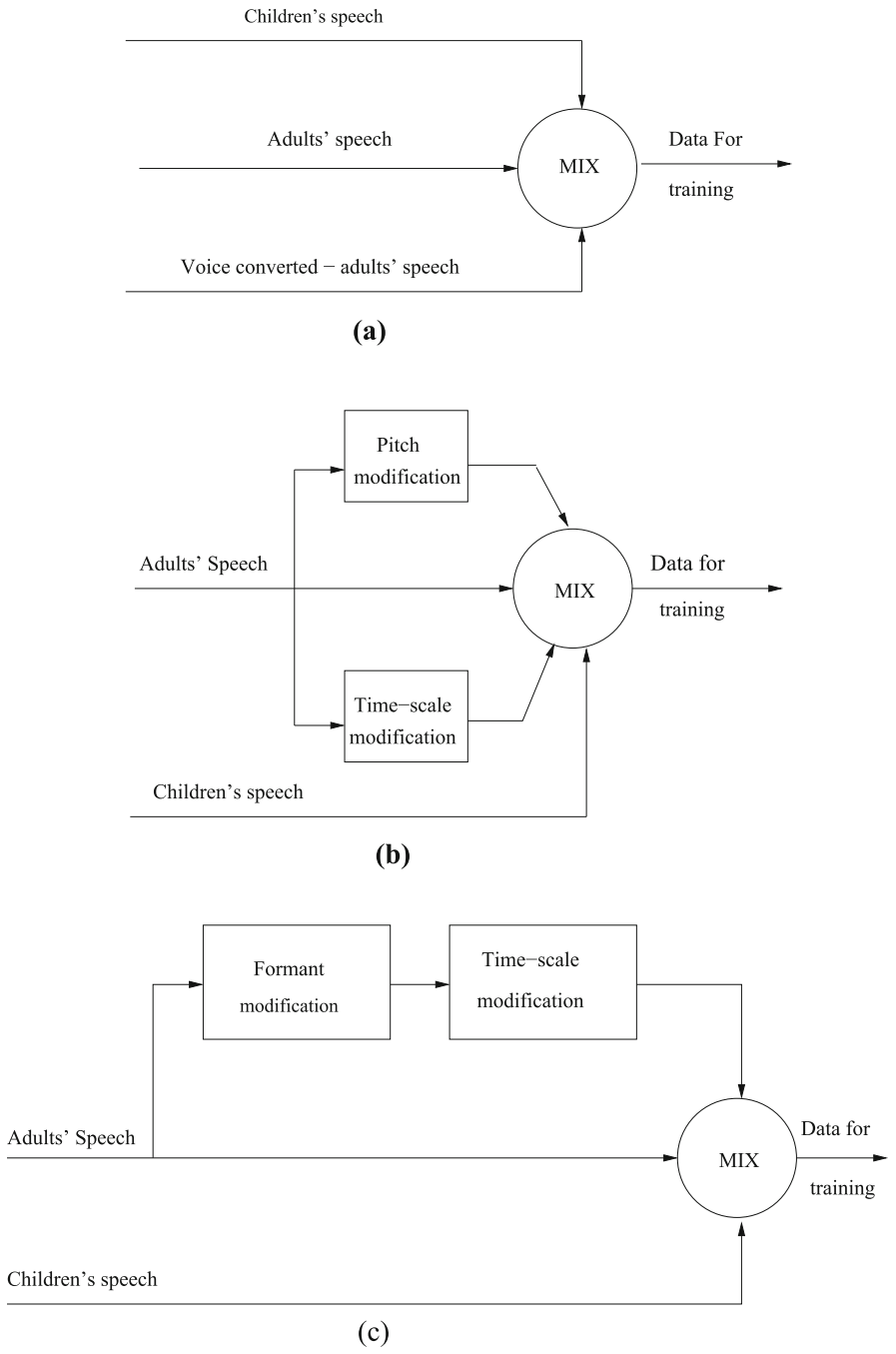

Fig. 1 Block diagram summarizing the different out-of-domain data augmentation techniques explored in this study

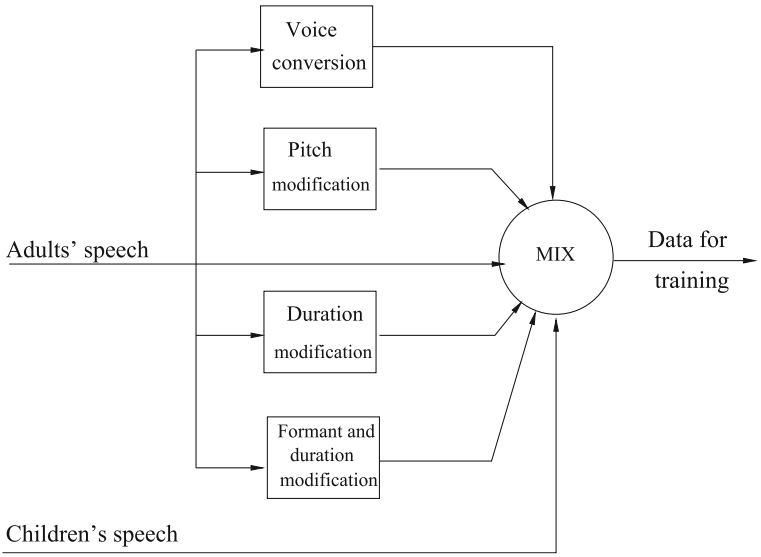

Fig. 2 Block diagram summarizing the out-of-domain data augmentation technique proposed in this paper

1. In the first approach, the adults' speech was subjected to voice conversion (VC) using a cycle-consistent generative adversarial network. The C-GAN was trained using 10 minutes of speech data from both adult as well as child speakers. Due to VC, adults' speech utterances sound very similar to children's speech as noticed during the listening tests. Finally, the voice converted adults' data was pooled with the children's speech as well as the original adults' speech. The model parameters were then trained on the pooled data. The overall scheme is pictorially represented in Fig. 1a.
2. In the second approach, children's speech was pooled together with adults' data along with the prosody modified (pitch and time-scale modification) adults' speech as shown in Fig. 1b. In this case, pitch of speech data from the adult speaker 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 the earlier reported works on children's speech recognition [23]. The motivation behind this approach lies in the fact that children's speech exhibits higher fundamental frequency or pitch as opposed to adult's speech. The variation of pitch among children of diversified ages is also very evident. Further, it is also a well-known fact that on an average, the speaking-rate for children is lesser than that for the adult speakers. Hence, stretching the speech utterances from adult speakers can compensate for the differences in the speaking-rates. In order to perform prosody modification (PM), the technique reported in [17] was used.
3. The formant frequencies in children's speech are higher as compared to those for the adults owing to the fact that vocal tract length is smaller in the case of child speakers. In addition to that the average phoneme duration in case of children is longer. As a result, the speaking-rate of children is slower as compared to adults

as already mentioned. These differences in the acoustic characteristics of a child speaker in contrast to an adult speaker leads us to the third approach for data augmentation wherein the formant frequencies (FM) of adults' speech data are 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) [17]. The modified adults' data was then pooled together with the children's speech data and the unperturbed adult speech data. This approach is pictorially represented in Fig. 1c. The mentioned scaling factors were determined from the earlier works as already mentioned [11, 23].

4. Finally, in the proposed approach of data augmentation, the speech data from children as well as the unperturbed adults' speech data were pooled together with all the modified versions of adults' data discussed above. The proposed data augmentation technique is summarized pictorially in the block diagram shown in Fig. 2. This helps to further increase the amount of training data. At the same time, all the targeted missing acoustic attributes are well-captured by the resulting training set.

It is noteworthy here that even though the aforementioned techniques of synthetically generating speech data are well-acclaimed in the literary works, their efficacy in the context of children's ASV systems using short utterances are relatively unexplored.

4 Role of feature concatenation

In the second part of this paper, as discussed earlier, we have studied the effect of concatenating two complementary front-end acoustic features at the frame-level in order to enhance the performance of the children's ASV system. In this regard, we have chosen the MFCC and FDLP features. The MFCC features are well-known and most commonly used front-end acoustic features in the context of speaker verification. The MFCC features model the spectral envelop corresponding to each of the short-time frames. However, the temporal structure is not effectively represented. In order to address this short-coming, the velocity and acceleration coefficients are generally appended to the base features. In recent years, time-splicing has been utilized in the place of appending velocity (delta) or acceleration (delta-delta) coefficients. This shows that effectively capturing the temporal envelop is also critical. Earlier works have shown that effective modeling of the temporal peaks can aid in improving the efficacy of several speech processing tasks [13, 31, 32]. The front-end acoustic features used in those works are FDLP features. The FDLP features capture the temporal envelope by applying linear predictive coding on the spectra rather than the time-domain representation. Hence, the MFCC and FDLP features represent complementary acoustic information. Motivated by the complementary nature of the two features, it is expected that frame-level concatenation of the two types of features will also enhance the performance of children's ASV task.

In Fig. 3, we have pictorially outlined the frame-level concatenation approach. Given the speech signal, first, we extract MFCC and FDLP features. Next, for each of the short-time frames, the corresponding MFCC and FDLP features are appended. The resulting feature vectors are used as the input to the *x*-vector extraction process

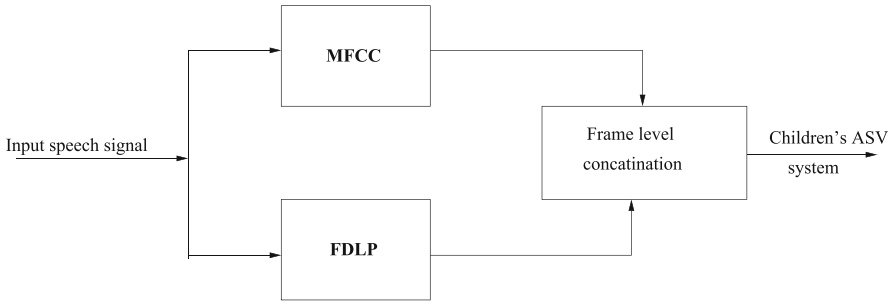

Fig. 3 Block diagram outlining the procedure for concatenating MFCC and FDLP features

instead of the MFCC features. It is worth highlighting here that the role of FDLP features in the context of children's ASV task is unexplored.

In order to study the effect of feature concatenation, the following analysis was performed. Three child speakers were chosen randomly from the available speech corpus. Next, we took the MFCC features corresponding to all the speech utterances from those speaker. Finally, t-SNE plot was drawn using the selected MFCC features, each speaker being treated as one class. The t-SNE plot corresponding to this study is shown in Fig. 4a. This study was repeated using FDLP features as well as after frame-level concatenation of MFCC and FDLP features. The corresponding t-SNE plots are shown in Fig. 4b and c, respectively. As evident from the t-SNE plots, the speaker clusters move farther apart when FDLP features employed in place of MFCC features. In addition to that overlap among the speaker clusters is significantly less when the two features are concatenated. Therefore, the proposed idea of frame-level concatenation is expected to enhance the discrimination among the speakers. The same has been experimentally verified in this paper.

Since the proposed feature concatenation approach involves MFCC and FDLP features, in the following subsections, the two types of features are discussed for the sake of completeness. The discussion given next closely follows the works reported in [3] and [1], respectively.

4.1 Mel-frequency cepstral coefficients (MFCC)

Mel-frequency cepstral coefficients are the most commonly used front-end acoustic features as already mentioned earlier. During the MFCC feature extraction process, the speech signal is first analyzed into overlapping frames of short duration followed by the computation of short-time Fourier transform (STFT). Next, spectral warping is done over a non-uniform frequency scale by using triangular Mel-filter-bank. Resultant power spectrum undergoes Logarithmic compression followed by discrete cosine transform (DCT). Applying DCT yields the real cepstrum (RC). The final feature vectors that are fed as input while training any classifier are obtained by low-time liftering of real cepstrum. The overall process is summarized in the block diagram shown in Fig. 5.

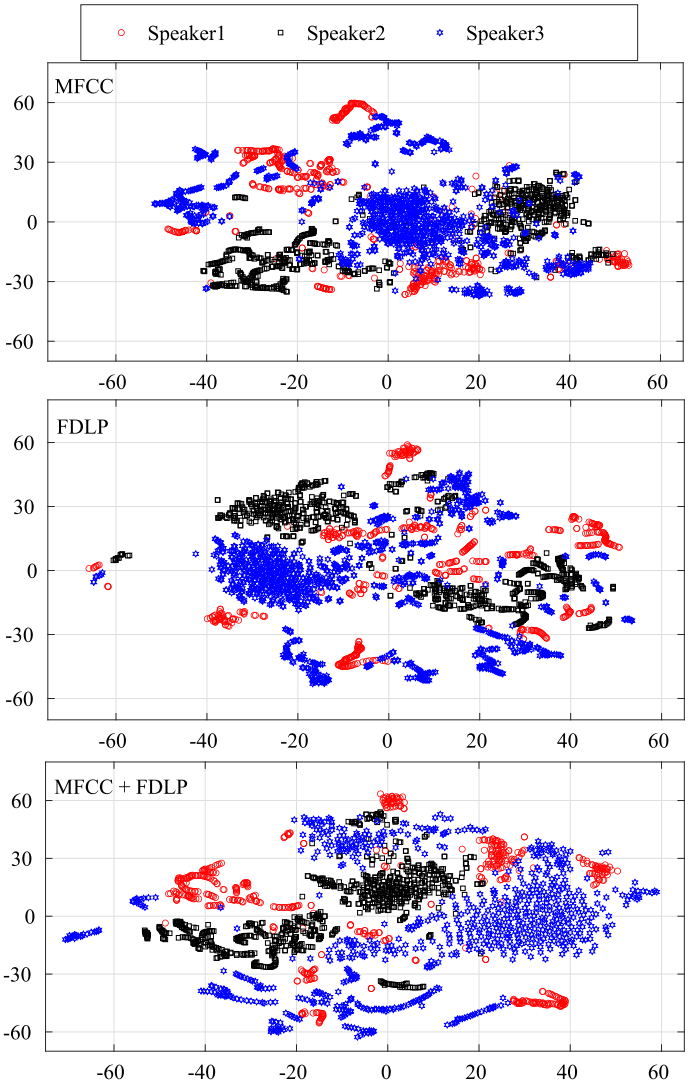

Fig. 4 t-SNE plots depicting the superior discrimination among speaker clusters obtained through frame-level concatenation of MFCC and FDLP features

4.2 Frequency-Domain Linear Prediction (FDLP) coefficients

Now, its time to redirect our focus to the extraction procedure of an alternative acoustic front-end feature utilized in this study, namely the FDLP. To facilitate the process of feature extraction, the input speech signal is divided into segments, each approximately 1000 milliseconds in duration. These segments are subsequently broken down into sub-bands, and the FDLP technique is employed to derive a parametric model that characterizes the temporal envelope. In the case of shorter utterances, the input

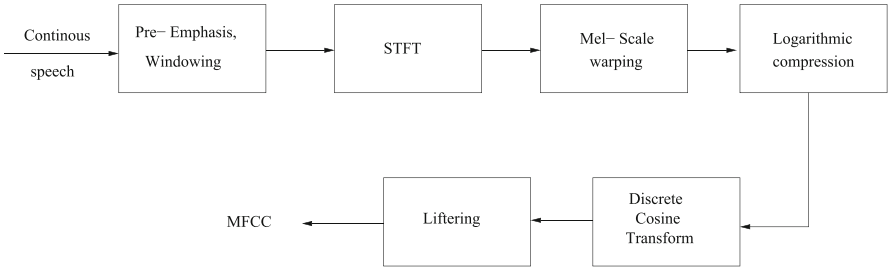

Fig. 5 Block diagram outlining the process of extracting MFCC features

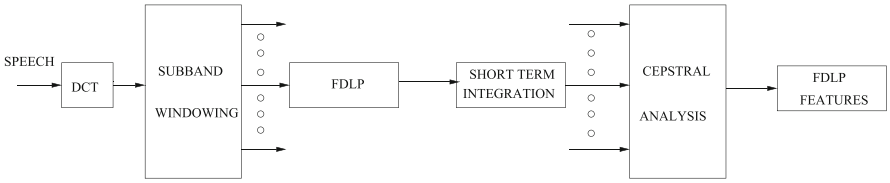

Fig. 6 Block diagram outlining the process of extracting FDLP features

signal is padded with zeros to ensure an adequate number of samples before sub-band decomposition. The collection of temporal envelopes from all the sub-bands creates a two-dimensional representation in the time-frequency domain for the input signal. This two-dimensional representation is subjected to convolution with a rectangular window having a duration of 25 milliseconds and a frame rate of 100 Hz, with 10 milliseconds of overlap between frames. These subsampled short-term spectral energies are subsequently transformed into short-term cepstral features. The entire process is succinctly illustrated using the block diagram depicted in Fig. 6

The FDLP feature extraction employed in the pilot study of this paper was carried out using three different filter-banks: Mel-, Bark- and linear filter-banks. The brief detail of each of these scales are summarized in the following: Mel-scale filter-banks are a set of triangular filters with a peak response equal to unity at the center frequency. The central frequency of each Mel-scale filter bank is uniformly spaced till 1000 Hz, and it follows a logarithmic scale thereafter. The mapping from linear frequency scale (f in Hz) to the Mel-frequency scale (m) is given by:

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

The Bark scale provides an alternative perceptually motivated scale to the Mel-scale. The basilar membrane (BM) which is an important part of the inner ear performs the spectral analysis followed by speech intelligibility perception in humans. Each point on the BM can be considered as a band-pass filter having a bandwidth equal to one critical bandwidth or one Bark. The bandwidth of several auditory filters were empirically observed and used to formulate the Bark scale. The transformation of

linear frequency scale(f in Hz) into Bark-frequency scale (B) [18, 29] is given by:

$$B = 13 \arctan\left(\frac{0.76f}{1000}\right) + 3.5 \arctan\left(\frac{f}{7500}\right)^2 \tag{2}$$

The linear filter-banks' resolution is alike for all the frequency components of the spectral information. The linear-scale frequency of linear filter-bank may prove to be beneficial for children's ASV system as the filter-bank coefficients cover all speech frequency ranges equally and considers them equally important. Earlier works have shown that the higher-frequency components in children's speech are richer in speaker-specific information. Hence, effectively preserving those by the use of linear filter-bank may improve the performance of an ASV system. The same has been experimentally verified in this study.

5 Experimental Evaluation

In this section, the relative effectiveness of the proposed data augmentation technique as well as the feature concatenation approach are experimentally verified and the results are presented.

5.1 The Speech Corpora

Four different speech corpora were utilized for the development and evaluation of children speaker verification system. Those corpora are CSLU kids corpus [27], CMU kids corpus [4], PF-STAR children's speech database [2] and WSJCAM0 adults' speech corpus [19]. The details of each datasets are as follows:

1. *CSLU kids corpus*: This dataset consisting of spontaneous and prompted speech comprising of 100 hours of data with 73, 100 utterances from 1100 children. The contributing children speakers are in-between kindergarten to grade 10. The speech data is sampled at 16 kHz rate. We have used this corpus as the training data in this work.
2. *CMU kids corpus*: This dataset comprises of 9.1 hours of data with 5, 180 utterances from 76 children. The child speakers belong to the of age group 6 to 11 years. This database is also sampled at 16 kHz rate. We have used this corpus as our test-set. A total of 423, 388 genuine and 26, 403, 832 impostor trails are present in this dataset. The average duration of the data in this corpus is 6seconds. Therefore, evaluation on this set represents the short-utterance case.
3. *WSJCAM0 corpus*: This adults' speech dataset is used for out-of-domain data augmentation. This corpus consists of 15.5 hours of data with 7, 852 utterances from 92 adult speakers (male and female). The sampling rate is 16 kHz. After performing several data modification techniques like voice conversion (VC), prosody modification (PM) and formant modification (FM) along with time scale modification (TSM), a total of 63 hours of synthetic data generated using WSJCAM0 for training purpose with acoustic attributes similar to those of children's speech.

4. *PF-STAR children's speech database*: This dataset comprising 8.3 hours of speech data from 121 speakers was utilized for the development set. A total of 6, 664 genuine trails and 995, 420 impostor trails are present in this development set. There are 856 utterances in this development set and duration of the utterances is very long (up to 3 minutes).

5.2 Performance Evaluation Metrics

To evaluate the performance of the developed short-utterance based children's ASV system, the metrics used in this study are: equal error rate (EER) and decision cost function (DCF). The EER is the location on the detection error trade-off (DET) plot where the likelihood of both false acceptance rate (FAR) and false rejection rate (FRR) is equal. Lower the value of EER, higher is the accuracy of the ASV system. The severity of the two types of errors may not be equivalent. Consequently, it makes sense to weight the two normalized error rates with

– The prior probability of targets in the expected application and
– The estimated costs of the two error types.

After applying these weightings, one obtains a scalar performance metric, popularly called as detection cost function (DCF). The DCF is computed using Eq. 3:

$$\mathbf{DCF} = (\mathbf{C_{FRR}} * \mathbf{FRR} * \mathbf{P_{Targ}}) + (\mathbf{C_{FAR}} * \mathbf{FAR} * (\mathbf{1} - \mathbf{P_{Targ}}) \tag{3}$$

where C_{FAR} and C_{FRR} stand for the cost of FAR and cost of FRR, respectively. P_{Targ} is defined as the prior probability that the test speech was made by the claimed speaker. The decision cost function for the NIST evaluation is represented as in Eq. 3, where P_{Targ} is set to 0.01, the cost of false alarm C_{FAR} is set to 1, and the cost of miss C_{FRR} is set to 10. The minimum value obtained on the test data is referred to as the minimum DCF (minDCF). It can be obtained by selecting the score threshold such that it minimizes Eq. 3 on the test data.

5.3 Experimental Setup

The entire set-up for the ASV system was developed using the kaldi toolkit [15]. As already stated earlier, we have used two front-end acoustic feature namely Mel-frequency cepstral coefficient and frequency-domain linear prediction coefficients. It is worth mentioning here that the MFCC features were extracted using the Kaldi toolkit while FDLP-based front-end speech parameterization was performed using MATLAB. In the process to extract those two kinds of front-end features, speech data were first high-pass filtered having pre-emphasis factor of 0.97. Each of the speech utterances were analyzed into short-time frames using overlapping Hamming windows. The duration of those overlapping Hamming windows were chosen to be 25 ms with a frame shift of 10 ms. A 30-channel log Mel-filter-bank was engaged for warping of spectrum before extracting 30-dimensional MFCC features. On the other hand, for the FDLP feature extraction, three different types of filter-banks, namely

Mel-scale, Bark-scale and linear-scale filter-bank, were employed. The number of modulation components and cepstral components were 14 and 30, respectively. The frame length were chosen to be 25 ms with a frame shift of 10 ms, similar to that of MFCC specifications.

The x-vector extractor consisted of a time-delay neural network (TDNN) architecture [28] comprising of 7 hidden layers trained for 6 epochs. The parameters of the network were trained using gradient descent algorithm [16, 28]. Each of the speech utterances was finally represented as a 512-dimensional x-vector.

5.4 Experimental Results and Discussions

For experimental evaluations, we first performed an initial study using the development set derived from the PF-STAR children's speech database. The training set was composed of the CMU kids corpus and the WSJCAM0 database along with its modified versions. Since the amount of data used for training was lesser in duration, the overall training time was also less. Hence, this study helped us to perform a larger number of experiments in lesser amount of time. It also helped us in reaching meaningful conclusions in lesser amount of time and then translate it to the case of short utterances. Furthermore, since the duration of the test utterances was significantly longer, it helped us in gauging the severity of the problem faced when the duration of the test data is reduced significantly. In the following, we first present the initial study employing the development set. This is followed by the experimental evaluations performed in the case short utterances.

5.4.1 An initial study using the development set

To test the feasibility, reliability and pertinence of the extracted front-end acoustic feature, FDLP individually as well as in tandem with the classical MFCC features, a pilot study was designed. The EER and minDCF values obtained for this pilot study evaluated against long utterances of children's speech test-set with respect to an ASV system trained on either children's speech or a mix of children's and adults' speech along with the modified adults' speech are given in Table 1. As evident from the table, the EER and minDCF undergo appreciable improvement with the implementation of the proposed data augmentation technique. A relative improvement of 35.86% with respect to the baseline ASV system trained exclusively on child dataset is achieved when the proposed data augmentation techniques is employed. This shows that the missing targeted attributes have been well-captured as the consequence of the proposed data augmentation. Consequently, the developed ASV system generalizes better for the children's speech. The separate impact of the augmented data for different setups are also enlisted in Table 1 for comparison with the data augmentation technique proposed in this paper.

Next, the effectiveness of the proposed feature concatenation approach was evaluated. The EER and minDCF values obtained when MFCC and FDLP features extracted from three different filter-banks were concatenated are given in Table 2 for the entire test-set. The EER and minDCF values obtained when the ASV system is trained

Table 1 EER and minDCF values for the children's long-utterance speech test set demonstrating the effectiveness of out-of-domain data augmentation techniques. The MODIFIED ADULT SPEECH in the out-of-domain data augmentation scheme includes adult voice conversion (ADULT-VC), adult time-scale modification (ADULT-TSM), adult formant modification (ADULT-FM) and adult pitch modification (ADULT-PM)

Data used for training	Evaluation Metric	
	EER(%)	minDCF
CHILD (Baseline)	4.235	0.6442
CHILD + ADULT	3.574	0.5432
CHILD + MODIFIED ADULT SPEECH	2.931	0.4565
CHILD + ADULT + MODIFIED ADULT SPEECH (***PROPOSED***)	**2.716**	**0.3655**

Bold used for specific numbers signify better performance of the proposed technique/approach in the paper, compared to the other techniques

either using only MFCC features or using only FDLP features are also enlisted for comparison. The proposed data augmentation technique has been employed prior to the training of the ASV system. As evident from Table 2, relevant improvements are observed in the evaluation metrics when MFCC features are concatenated with the FDLP features. The frame-level concatenation of the MFCC features with the FDLP features extracted after employing the linear filter-bank out-classes all other feature concatenation pairs and culminates in a significant relative improvement of 22.34% in EER on the entire test-set.

The preliminary study discussed so far was conducted on a relatively smaller sample size but serves as a valuable preparatory step before the main rigorous research process is undertaken. The takeaways of this pilot study are as follows: Firstly, the evaluation metrics obtained by the implementation of FDLP features alone is only slightly better as compared to the implementation of only MFCC features. This suggests that modeling temporal envelopes alone (by using FDLP) or modeling spectral envelopes alone (by using MFCC) is not enough. Concatenation of MFCC with FDLP, which simultaneously does spectral as well as temporal estimation, leads to more effective results in terms of EER and minDCF. Secondly, it has also made evident that the feature fusion of MFCC features with the FDLP features extracted from linear filter-banks provides the most desirable results in context to children's ASV system. The linear-scale frequency of linear filter-bank proves to be beneficial for children's ASV system as compared to Mel-scale and Bark-scale since it effectively preserves the higher-frequency components in children's speech which are richer in speaker-specific information.

5.4.2 Rigorous Study: Case of Short utterances

In the previous subsection, the development set comprising of very long utterances from children emphasized the efficacy of the frame level concatenation of MFCC features with the FDLP features. Taking a cue from this, an extensive study involving feature fusion model of MFCC and FDLP coefficients (extracted from linear filter-banks) is undertaken to train the children's ASV system on a large training dataset and

Table 2 EER and minDCF values for the ASV system trained on the full dataset obtained using the proposed data augmentation technique, demonstrating the effectiveness of feature concatenation. This study was performed on *x*-vector-based children's ASV system employing long utterances for testing

Features	EER (%)	minDCF
MFCC	2.716	0.3655
FDLP_Linear	2.412	0.3588
MFCC + FDLP_Bark	2.395	0.3448
MFCC + FDLP_Mel	2.287	0.3631
MFCC + FDLP_Linear	**2.109**	**0.3313**

Bold used for specific numbers signify better performance of the proposed technique/approach in the paper, compared to the other techniques

evaluated on short-utterances of children's test-set. In this case, the CSLU kids corpus was used for training purpose while the CMU kids database was used for evaluation. The details of the different datasets used for training and testing purposes for both the development set as well as the evaluation set is shown in Table 3.

The EER and minDCF values for the children's short-utterance test set with respect to an ASV system trained on either children's speech or a mix of children's and adults' speech along with the modified adults' speech are given in Table 4. When the proposed out-of-domain data augmentation is applied, the EER of the ASV system climbs from 2.716% (for long utterances of development set) to 14.58% (for short utterances of test-set), despite the fact that the amount of training data used in the latter case is 178.5 hrs as opposed to 87.6 hrs of training data in the former case. This brings forth the gravity of the task in hand, namely the short-utterance case. Again, as evident from the Table 4, the EER and minDCF values undergo successive improvement with the application of subsequent data augmentation technique. A relative improvement of 33.6% with respect to the baseline system trained on child dataset alone is achieved when the proposed data augmentation techniques is employed.

Next, we evaluated the effectiveness of the proposed feature concatenation approach. The EER and minDCF values obtained when MFCC and FDLP features are concatenated are given in Table 5. In this 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 are also enlisted for comparison. As evident, an absolute reduction by 1% is achieved by feature concatenation. The detection error trade-off (DET) plot summarizing these results is shown in Fig. 7. In this plot, baseline refers to the ASV trained exclusively on children's speech using MFCC features.

To gauge the effectiveness of proposed approaches in more detail, we have also performed an age-wise analysis as well as gender-wise analysis of children's speech. To evaluate the effect of age variation, the test-set was split into two groups. The first consisted of data from speakers belonging to age-group $6-7$ years while the second one comprised of speech utterances from speakers belonging to age-group $8-9$ years. The EER and minDCF values for this experimental study are given in Table 6. In this case as well, the proposed out-of-domain data augmentation has been

Table 3 Details of the different datasets used for training and testing

Type of Experimental Set	Corpus	Duration of Data (in Hrs)	Number of Speakers	Split
Development Set	CHILD(CMU kids corpus)	9.1	76	Train
(Long Utterances)	ADULT(WSJCAM0 corpus)	15.5	92	Train
	ADULT-VC	15.5	92	Train
	ADULT-PM	31.5	184	Train
	ADULT-FM-TSM	16	92	Train
	Total (Children-Train)	**87.6**	**536**	Train
	CHILD(PF-STAR kids corpus)	8.3	121	Test
Evaluation Set	CHILD(CSLU kids corpus)	100	1100	Train
(Short Utterances)	ADULT(WSJCAM0 corpus)	15.5	92	Train
	ADULT-VC	15.5	92	Train
	ADULT-PM	31.5	184	Train
	ADULT-FM-TSM	16	92	Train
	Total (Children-Train)	**178.5**	**1560**	Train
	CHILD(CMU kids corpus)	9.1	76	Test

Bold used for specific numbers signify better performance of the proposed technique/approach in the paper, compared to the other techniques

Table 4 EER and minDCF values for the children's short-utterance speech test-set with respect to an ASV system trained on either children's speech or a mix of children's and adults' speech along with the modified adults' speech

Dataset	Evaluation metric	
	EER (%)	minDCF (%)
CHILD	21.95	0.9975
CHILD + ADULT-FM-TSM	19.78	0.9881
CHILD + ADULT-VC	17.34	0.9751
CHILD + ADULT-PM	16.30	0.9492
CHILD + ADULT + ADULT-FM-TSM + ADULT-PM + ADULT-VC (***PROPOSED***)	**14.58**	**0.9233**

Bold used for specific numbers signify better performance of the proposed technique/approach in the paper, compared to the other techniques

Table 5 EER and minDCF values with respect to the ASV system trained on the dataset obtained using the proposed out-of-domain data augmentation technique and subjected to children's short-utterances, demonstrating the effectiveness of feature concatenation

Acoustic features	Evaluation metric	
	EER (%)	minDCF
MFCC	14.58	0.9233
MFCC + FDLP	**13.60**	**0.9014**

Bold used for specific numbers signify better performance of the proposed technique/approach in the paper, compared to the other techniques

employed before training the *x*-vector extractor. As can be seen from the listed results in Table 6, a significant degradation (reflected by 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 may be attributed to various factors, such as the development of language skills, motor control, and cognitive processing as children grow and mature. Due to the inherent shorter vocal tracts, children's speech has higher formant frequencies and pitch frequencies, which contribute to the degradation. The speaking rate tends to stabilize as children grow and the formant frequencies subsides. Further as is evident from Table 6, the ASV system trained solely on the MFCC features perform poorly in terms of evaluation metrics. 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 FDLP features with the MFCC features represent the spectral as well as the temporal structure effectively.

The EER for the full test-set shows a relative reduction of 6.72% on the frame-level concatenation of MFCC and FDLP feature, pictorially depicted by the first bar in the Fig. 8. The corresponding relative reduction in EER for the age group $6-7$ years is **11.90%**, pictorially represented by the second bar in Fig. 8. The EER of age

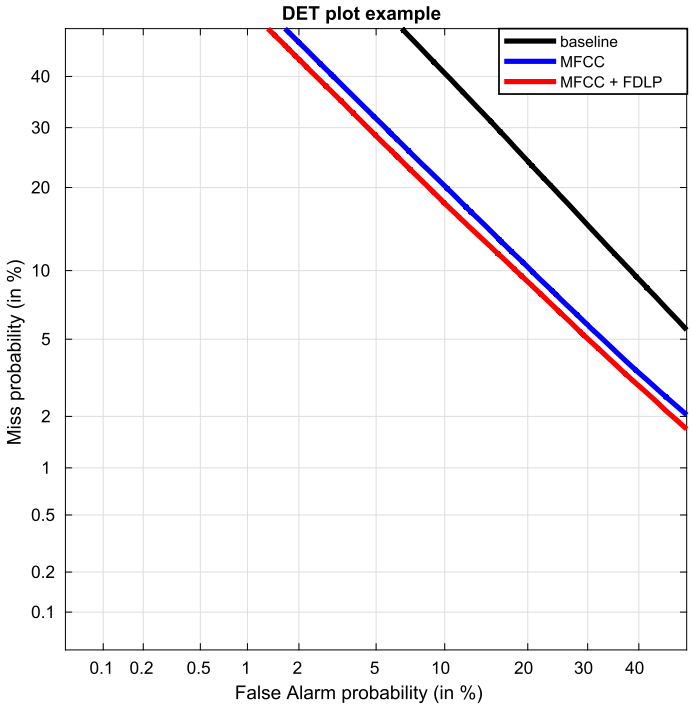

Fig. 7 Detection error trade-off plot demonstrating the effectiveness of proposed feature concatenation approach

group $8-9$ years shows a relative reduction of 9.25% when MFCC and FDLP are concatenated, pictorially depicted by the third bar in Fig. 8.

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 7. 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 and pitch of female children's speech compared to their male counterparts. Further, as evident from the table, the EER considerably reduces when FDLP features are concatenated with MFCC features. The relative reduction in EER for the female child on the frame-level concatenation of the features is 9.27%, pictorially depicted by the fourth bar in Fig. 8. For the male child, the relative reduction in EER is 8.55% as compared to the baseline, pictorially depicted by the fifth bar in Fig. 8.

Table 6 Age-group-specific EER and minDCF values with respect to the ASV system trained on the dataset obtained using the proposed out-of-domain data augmentation technique

Acoustic features	Age group	Evaluation metric EER (%)	minDCF
MFCC	6-7	17.39	0.9657
	8-9	14.04	0.9203
MFCC + FDLP	6-7	**15.32**	**0.9420**
	8-9	**12.74**	**0.8983**

Bold used for specific numbers signify better performance of the proposed technique/approach in the paper, compared to the other techniques

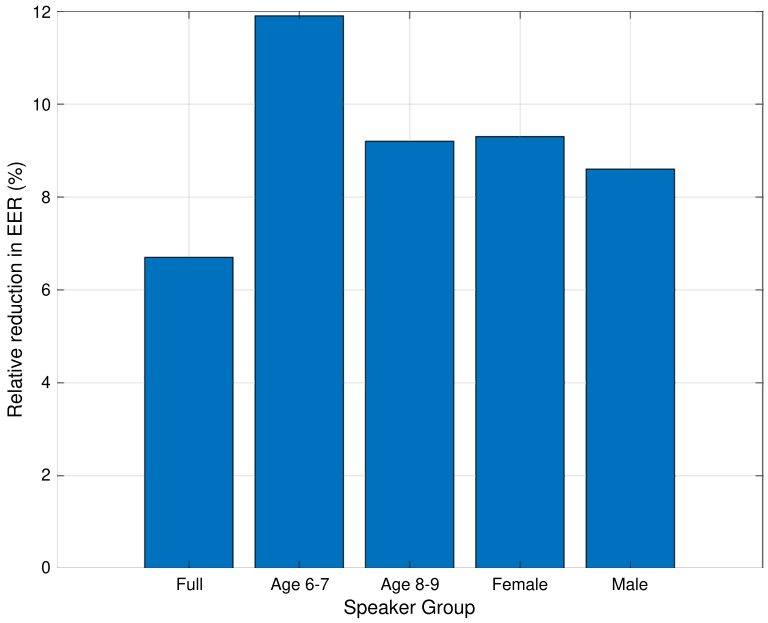

Fig. 8 Bar graph representation of the relative reduction in EER(%) for various speaker groups(in terms of full test-set, age and gender) corresponding to the ASV system trained on the concatenated MFCC and FDLP features

6 Summarized Overview

To sum it up, an overview of the discussed approaches in this paper along with their advantages and limitations are summarized in Table 8.

7 Future Scope

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

Table 7 Gender-wise break up of EER and minimum DCF values highlighting the significance of feature concatenation approaches. 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

Acoustic features	Child Gender	Evaluation metric	
		EER (%)	minDCF
MFCC	Female	17.68	0.9585
	Male	10.05	0.8534
MFCC+FDLP	Female	16.04	0.9382
	Male	**9.19**	0.8049

Bold used for specific numbers signify better performance of the proposed technique/approach in the paper, compared to the other techniques

Table 8 Overview of the discussed approaches in the implementation of short-utterance-based children ASV system

Discussed Approaches	Advantages	Limitations
1. Out-of-domain data augmentation, employing speech processing techniques to modify the acoustic attributes of adults' data to match children's speech. These techniques include VC, PM, FM-TSM	– Increase in amount of training data leading to more robust estimation of model parameters; – Modifying the acoustic attributes of adult ensures that the ASV system does not get biased toward adult speakers.	– Finding an optimal data augmentation techniques for the data training is non-trivial; – The underlying bias of the original data is maintained in the augmented data; – Implementation of VC using C-GANs is computationally quite intensive.
2. Front-end feature extraction: MFCC	– Traditional front-end acoustic feature which effectively models the spectral envelop at segmental(10-30 ms) levels of speech; – They provide compact and stable representation of vocal tract of a speaker	– They fail to represent details at sub-segmental (3-5 ms) and supra-segmental (100-300 ms) levels of speech. Temporal structures are only weakly represented, even after employing time-splicing techniques.
3. Frame level feature concatenation of MFCC with the FDLP features	– MFCC and the FDLP features containing complementary acoustic information which helps in simultaneously modeling the spectral as well as temporal envelopes; – Proposed feature concatenation results in speaker clusters to move significantly apart, enhancing discrimination among the speakers.	– Increase in dimensionality due to feature concatenation leads to an increase in computational complexity

acoustic attributes of children's speech for training. In particular, we would 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 dataset 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. In addition, the works discussed in [30, 34] were found highly influential, their findings as vital and the authors' would like to integrate these findings into children's speaker verification tasks in their future endeavor. In [30], a robust algorithm is used for identifying output error (OE) models with constrained output in the presence of non-Gaussian noises, addressing practical challenges such as rare, inconsistent observations and outliers. The algorithm, based on Huber's robust statistics theory, considers the significant role of constraints in ensuring control performance and process safety. The proposed robust algorithm, enhanced by optimal input design using a minimum variance controller, demonstrates improved accuracy in parameter estimates for OE models compared to linear identification algorithms, with simulations illustrating enhanced convergence rates. The study in [34] focuses on addressing the issue of hybrid-driven fuzzy filtering for nonlinear semi-linear parabolic partial differential equation systems facing dual cyber attacks, including deception and denial of service attacks. The approach involves employing a Takagi-Sugeno fuzzy model for system reconstruction, applying a hybrid-driven mechanism for filter design to balance system performance and limited network resource consumption, and using the Lyapunov direct method to establish stability conditions for the augmented system.

8 Conclusion

The work in this paper outlines our efforts to create a reliable children's ASV system employing short-utterances under low-resource conditions. To address the inevitable problem of speech data paucity, various out-of-domain data augmentation techniques were explored to synthetically generate more data for training. Interestingly, all the augmentation techniques have shown improvement from the previous augmented data as evident from the result section. When the proposed data augmentation approach is used, a relative improvement of 33.57% is made compared to the baseline system trained on the child dataset alone. 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. In addition to data augmentation, the effectiveness of frame-level concatenation of MFCC with the FDLP, is also examined in this paper. Traditional front-end acoustic features such as MFCCs model the spectral envelop corresponding to each of the short-time frames. On the other hand temporal structures are only weakly represented, even after employing time-splicing techniques. The frame-level concatenation of the MFCC features with the FDLP features are demonstrated to simultaneously model the spectral as well as temporal envelopes in this work. 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 exhibiting more pronounced inter-speaker variability, results 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 as well as feature concatenation is found to be more effective for children in the lower age-group. The findings of this study will provide a foundation for future advancements in children speaker verification systems in the context of short utterances and contribute toward improved security, personalized experiences, and educational opportunities for children while ensuring their safety and well-being.

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

Data Availability The datasets 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. M. Athineos, D. Ellis: Frequency-domain linear prediction for temporal features. In: 2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721), pp. 261–266 (2003)
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: Proc. INTERSPEECH, pp. 2761–2764 (2005)
3. S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoustic, Speech Signal Processing **28**(4), 357–366 (1980). https://doi.org/10.1109/TASSP.1980.1163420
4. M. Eskenazi, J. Mostow, D. Graff: The CMU Kids Corpus LDC97S63. https://catalog.ldc.upenn.edu/LDC97S63 (1997)
5. S. Fernando, V. Sethu, E. Ambikairajah: Sub-band envelope features using frequency domain linear prediction for short duration language identification. In: INTERSPEECH, pp. 1818–1822 (2018)
6. S. Ganapathy, P. Rajan, H. Hermansky: Multi-layer perceptron based speech activity detection for speaker verification. In: 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 321–324. IEEE (2011)
7. M. Islam: Frequency domain linear prediction-based robust text-dependent speaker identification. In: 2016 international conference on innovations in science, engineering and technology (ICISET), pp. 1–4. IEEE (2016)
8. T. Kaneko, H. Kameoka: Parallel-data-free voice conversion using cycle-consistent adversarial networks. arXiv preprint arXiv:1711.11293 (2017)

9. H.K. Kathania, S.R. Kadiri, P. Alku, M. Kurimo: Study of formant modification for children asr. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 7429–7433. IEEE (2020)
10. R. Kethireddy, S.R. Kadiri, S.V. Gangashetty, Exploration of temporal dynamics of frequency domain linear prediction cepstral coefficients for dialect classification. Appl. Acoustics **188**, 108553 (2022)
11. 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)
12. S. Lee, A. Potamianos, S. Narayanan, Acoustics of children's speech: Developmental changes of temporal and spectral parameters. J. Acoust. Soc. Am. **105**(3), 1455–1468 (1999)
13. G. Mantena, S. Achanta, K. Prahallad, Query-by-example spoken term detection using frequency domain linear prediction and non-segmental dynamic time warping. IEEE/ACM Trans. Audio, Speech, nd Language Process. **22**(5), 946–955 (2014)
14. A. Poddar, M. Sahidullah, G. Saha, Speaker verification with short utterances: a review of challenges, trends and opportunities. IET Biometrics **7**(2), 91–101 (2018)
15. 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: Proc. ASRU (2011)
16. D. Povey, X. Zhang, S. Khudanpur: Parallel training of dnns with natural gradient and parameter averaging. arXiv preprint arXiv:1410.7455 (2014)
17. S.R.M. Prasanna, D. Govind, K. S. Rao, B. Yegnanarayana: Fast prosody modification using instants of significant excitation. In: Proc. Int. Conf. on Speech Prosody (2010)
18. T. F. Quateier: Discrete time processing of speech signals- principles and practice (1997)
19. T. Robinson, J. Fransen, D. Pye, J. Foote, S. Renals: WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition. In: Proc. ICASSP, vol. 1, pp. 81–84 (1995). https://doi.org/10.1109/ICASSP.1995.479278
20. S. Safavi, M. Russell, P. Jančovič, Automatic speaker, age-group and gender identification from children's speech. Comput. Speech & Language **50**, 141–156 (2018)
21. M. Sahidullah, T. Kinnunen, C. Hanilçi: A comparison of features for synthetic speech detection (2015)
22. S. Shahnawazuddin, N. Adiga, H.K. Kathania, B.T. Sai, Creating speaker independent asr system through prosody modification based data augmentation. Pattern Recognition Lett. **131**, 213–218 (2020)
23. 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
24. 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. Digital Signal Processing **93**, 34–42 (2019)
25. 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: Proc. ICASSP, pp. 7554–7558 (2020). https://doi.org/10.1109/ICASSP40776.2020.9053891
26. S. Shahnawazuddin, W. Ahmad, N. Adiga, A. Kumar, Children's speaker verification in low and zero resource conditions. Digital Signal Processing **116**, 103115 (2021)
27. K. Shobaki, J.P. Hosom, R. Cole: Cslu: Kids' speech version 1.1. Linguistic Data Consortium (2007)
28. D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur: X-vectors: Robust dnn embeddings for speaker recognition. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 5329–5333. IEEE (2018)
29. S.S. Stevens, J. Volkmann, E.B. Newman, A scale for the measurement of the psychological magnitude pitch. J. Acoustical Soc. Am. **8**(3), 185–190 (1937)
30. V. Stojanovic, N. Nedic, Robust identification of oe model with constrained output using optimal input design. J. Franklin Inst. **353**(2), 576–593 (2016)
31. S. Thomas, S. Ganapathy, H. Hermansky, Recognition of reverberant speech using frequency domain linear prediction. IEEE Signal Process. Lett. **15**, 681–684 (2008)
32. B. Wickramasinghe, S. Irtza, E. Ambikairajah, J. Epps: Frequency domain linear prediction features for replay spoofing attack detection. In: Interspeech, pp. 661–665 (2018)
33. G. Yeung, A. Alwan: On the difficulties of automatic speech recognition for kindergarten-aged children. Interspeech 2018 (2018)

34. Z. Zhang, X. Song, X. Sun, V. Stojanovic, Hybrid-driven-based fuzzy secure filtering for nonlinear parabolic partial differential equation systems with cyber attacks. Int. J. Adapt. Control Signal Process. **37**(2), 380–398 (2023)