Circuits, Systems, and Signal Processing (2024) 43:3139–3159
https://doi.org/10.1007/s00034-024-02598-1

# Role of Data Augmentation and Effective Conservation of High-Frequency Contents in the Context Children's Speaker Verification System

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

Received: 1 June 2023 / Revised: 25 December 2023 / Accepted: 26 December 2023 /
Published online: 5 February 2024


## Abstract

Developing an automatic speaker verification (ASV) system for children's speech presents significant challenges. One major obstacle is the scarcity of domain-specific data. This issue is exacerbated when dealing with short speech utterances, a relatively unexplored area in children's ASV. Voice biometric systems struggle during enrollment and verification phase, when faced with inadequate speech data, both in volume as well as in duration. To address data scarcity, this paper explores various in-domain and out-of-domain data augmentation techniques. Out-of-domain data from adult speakers, which have distinct acoustic attributes from children, are modified using techniques like voice-conversion, prosody and formant modification to make them acoustically similar to children's speech. In-domain data augmentation involves perturbing the speed of children's speech. This combined data augmentation approach not only increases training data volume but also captures missing target attributes, resulting in a significant 43.91% reduction in equal error rate (EER) compared to the baseline system. Additionally, the paper addresses the challenge of preserving higher-frequency components in children's speech. It achieves this by concatenating conventional Mel-frequency cepstral coefficients (MFCC) with Inverse-Mel-frequency cepstral coefficient (IMFCC) features at the frame level. The low canonical correlation between MFCC and IMFCC feature vectors motivates this fusion. The feature concatenation approach, when combined with proposed data augmentation, results in an appreciable reduction of 48.51% in the overall EER, demonstrating its effectiveness in improving the performance of children's ASV system.

✉ 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

**Keywords** Automatic speaker verification · In-domain data augmentation · Out-of-domain data augmentation · Mel-frequency cepstral coefficients · Inverse-Mel-frequency cepstral coefficients · Feature concatenation

## 1 Introduction

The web of cybernated applications in this digital age has fascinated people cutting across generations. Multimedia technologies and the Internet are fast evolving and has brought the whole world at our fingertip. In addition to the galore of positive aspects of cybernated applications is its dreary face. It is also fraught with the dangers of losing sensitive data and identity theft, if not accessed with caution. People accessing the online tools should be mindful of the cyber crimes and cyber frauds. To address such an intimidating issue, the field of biometrics have witnessed a meteoric rise in the recent past and is bound to remain at the center stage in the times to come. Voice/ Speech signal is one such biometric, which falls under the category of behavioral biometrics [9]. Even though the primary function of speech signal is human communication, it also captures information about the speaker's identity, age, emotions, gender, geographical origin and health. Voice biometrics or Automatic Speaker Verification (ASV) is a technology that uses algorithms and machine learning techniques to verify the identity of a speaker based on their speech characteristics. It is a biometric authentication method that relies on the unique patterns and traits in an individual's speech.

The process of speaker verification typically involves four main stages: enrollment, training, verification and decision making.

- Enrollment: During the enrollment phase, the system captures the speaker's voice samples typically through a microphone and extracts relevant acoustic features that are unique to their speech. These features can include aspects, such as pitch, frequency, duration, and spectral characteristics. The system then creates a speaker model or a template or a unique voiceprint based on these extracted features, which serves as a reference for future verification.
- Training: The extracted acoustic features are then used to train the ASV system. Machine learning algorithms analyze the acoustic features in the voiceprints and create a statistical model that captures the speaker-specific patterns. This training process helps the system learn to distinguish between different speakers and identify the unique characteristics of each individual.
- Verification: Once the enrollment and training stages are complete, the ASV system is ready for verification. In this stage the system compares the test speech sample of an individual to the stored speaker models. The incoming test speech sample is processed to extract similar features as those used during enrollment. The system then applies pattern matching algorithms and statistical models to compare the extracted features with the reference speaker models.
- Decision Making: The system calculates the similarity between the features of the test speech signal and the stored voiceprints of enrolled speakers. If the similarity score exceeds a predetermined threshold, the speaker is verified as the claimed identity; otherwise, the verification fails and the system rejects the user's claim.

As compared to other competing biometrics, voice biometrics is increasingly becoming popular because of its low cost, ease of use, faster authentication process and higher level of security features [9]. These striking features have caught the attention of researchers across the globe and considerable work has been reported on the development of ASV systems. But, the majority of the work reported in the literature deal with the design and development of ASV systems for adults. The fact that social networking websites and online learning tools are a rage among children and teenagers, with over half of youngsters in the age bracket of 6-15 obsessively indulging in internet and maintaining accounts on social media websites [2], cannot be denied. The children who are oblivious of the lurking perils in the usage of cyber related activities are the more vulnerable lot as opposed to the adults. This calls for the need of a robust ASV system for children. The literary works reported on building an ASV system for children are not vast as compared to adults [26, 29, 35]. Motivated by this, the authors' in this paper have focussed their attention on developing robust ASV systems for child speakers.

Modern ASV systems are found to be highly effective, resulting in a nominal error when supplied a sufficiently larger quantity and longer duration of speech data. State-of-the-art ASV systems employ deep learning architectures that necessitate estimation of a vast number of parameters. This, in turn, mandates a substantial quantity of domain-specific data. The road along the development of a reliable children's ASV system has many hindrances. The majority of children's speech corpora are not readily accessible. Moreover, these are limited in terms of data hours and the number of languages in which they are available. Developing an ASV system for languages without any children's speech corpus (zero-resource condition) is very demanding. Even if a small quantity of children's speech data is available (low-resource condition), designing an effective ASV system for children using deep learning architectures is still a very challenging task. Some of the earlier works on children's ASV have investigated the effect of synthetically generating speech data and then pooling it for training in order to circumvent the problem posed by low- and zero-resource conditions. It has been reported that out-of-domain data augmentation and in-domain data augmentation is effective in this regard [29]. The performance of an ASV system for children is further dented when there is a reduction in the duration of the speech utterances during testing, commonly termed as short-utterance situation. Speech segments of duration 5-10 seconds are commonly termed as short-utterances in the literary domain [10, 18]. In the context of an ASV system, it is observed that these systems show a decline in their performance as a consequence of reduction in the amount of speech, either during enrollment or verification stage [19, 34]. The primary challenge in achieving better results with short-utterances is the rise in intra-speaker variability of estimated parameters. Short utterances exhibits greater variability, which diminishes as the duration of utterances increases. The performance of ASV systems significantly deteriorates when the duration of speech is reduced, primarily because short utterances lack sufficient information to support accurate verification. The unavailability of sufficiently longer duration of speech data can be tackled during training phase by some data augmentation techniques. However, it is not feasible to do the same during the testing phase [18]. The works reported on children's ASV hardly deal with such short utterances scenario. In [30], the authors' had developed a children ASV system employing

in-domain and out-of-domain data augmentation techniques. However, the effect of formant modification of the adults' speech data was not studied in that paper. Moreover, the acoustic feature explored in that work was the conventional MFCC features alone. In addition, the performance of the developed ASV system was not evaluated on short-utterances of children's speech. In [1], the authors have looked for solutions beyond the classical MFCC features and have proposed the feature concatenation approach to preserve the higher-frequency components in children's speech. But [1], have studied and implemented only out-of-domain data augmentation techniques, it does not explore the scope of any in-domain data augmentation technique to minimize the equal error rate (EER) of the developed ASV system.

Taking cognizance of the above literary gap, the authors' have explored the role of both the in-domain as well as out-of-domain data augmentation techniques in order to synthetically generate speech more data. The goal of data augmentation is to artificially increase the size and diversity of the training data-set by applying various transformations to the original speech data. This technique helps in improving the generalization and robustness of the ASV system. The in-domain data augmentation technique used in this paper includes the default three-way speed perturbation of the original children's speech using Kaldi pipeline. To address the paucity of the domain-specific data, the impact of out-of-domain data augmentation techniques in the light of short-utterance-based children's ASV system is also explored in this paper. This includes (i) voice conversion (VC) of adults' speech data through a cycle-consistent generative adversarial network (C-GAN) [11], (ii) prosody modification (PM) [27, 28] 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) [12, 15] of adults' speech data. All the explored techniques not only help in increasing the amount of training data but also in modifying the acoustic attributes of adult's speech so that the acoustic mismatch with child's speech is minimal. The proposed combination of in-domain and out-of-domain data augmentation technique is observed to be highly effective as is demonstrated and validated in the experimental evaluation section in this paper.

Besides data augmentation, this exploration also delves into the role of feature concatenation of two front end acoustic features namely the Mel-frequency cepstral coefficients (MFCC) and the inverse-Mel-frequency cepstral coefficients (IMFCC). In general, the Mel-frequency cepstral coefficients (MFCC) are the most commonly used front-end acoustic features and have been popular ever since its inception. They provide a compact and stable representation of the vocal-tract of a speaker, which can capture speaker-specific characteristics. The MFCC features are extracted by projecting the power spectra onto Mel-weighted filter-banks. The configuration of this filter-bank involves a set of nonlinearly placed triangular filters, with each successive filters having bandwidth greater than the filter preceding it. When it comes to children's speech, a significant amount of relevant information is predominantly present in the high-frequency region [5, 25]. As resolution of Mel-filter-bank decreases with increase in frequency, the performance of children's ASV system based solely on MFCC features will be sub-optimal. In order to effectively preserve the higher-frequency contents in children's speech, the other front end acoustic feature explored in this paper is IMFCC. The IMFCC features are extracted using the inverse-Mel filter-banks, which

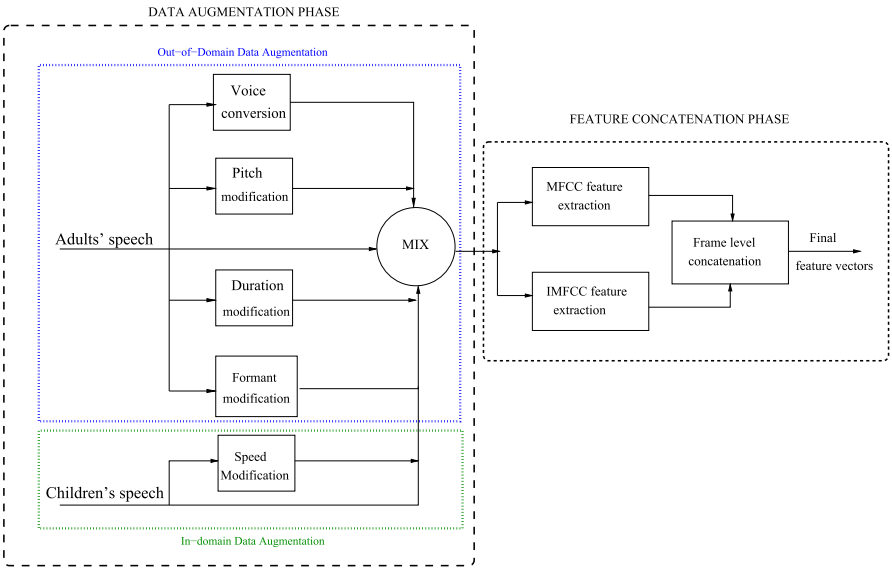


**Fig. 1** Block diagram outlining the data augmentation and feature concatenation approaches proposed in this work in order to enhance the verification performance of a short-utterance-based children ASV system

in turn are obtained by simply flipping the Mel-filter-banks around the midpoint. The inverse-Mel filter-banks is thus supposed to have better frequency resolution in higher-frequency range while the lower frequency components are down-sampled. As already highlighted, the use of Mel-filter-bank down-samples the spectral information in the higher-frequency range. The IMFCC due to its complementary nature of the filter-bank are supposed to better capture the acoustic information in the higher-frequency regions of children's speech, which are otherwise disregarded by the MFCC features. The feature fusion model of MFCC and IMFCC is thus expected to outperform the traditional MFCC, leading to an enhanced performance of children's ASV system.

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's ASV performance to unravel the effect of data augmentation and feature concatenation. One of the metrics used for the performance evaluation of the employed ASV system is Equal Error Rate (EER), which quantifies the system's ability to simultaneously balance false acceptance (verifying an impostor) and false rejection (rejecting a legitimate speaker) rates. Lower the value of EER, higher is the accuracy of the ASV system. The proposed approach also aids in diminishing the other evaluation metric called Detection Cost Function (DCF) considerably as opposed to the baseline system trained exclusively on children's speech alone using MFCC features. The ASV system 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 rest of this paper is organized as follows: Sect. 2 deals with an extensive exploration of in-domain data augmentation and various out-of-domain data augmentation techniques to deal with the scarcity of domain-specific data. In Sect. 2.3, we have talked about the authors' motivation to look beyond the traditional Mel-based filterbank and delve into the scope of feature concatenation for children's ASV system. The experimental evaluations exhibiting the efficacy of our proposed techniques are presented in Sect. 4. Eventually, conclusion is drawn in Sect. 5.

## 2 Explored Data Augmentation Techniques

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) [17, 32, 33] is trained. Deep learning models such as a TDNN have an inherent complexity owing to a number of hidden layers and hidden nodes per layer. They are resource intensive and require massive amount of data. The *x*-vectors are reported to be highly effective when the training data is in abundance. As already mentioned, one of the hindrances in the development of a reliable ASV system for children is the paucity of domain-specific data. Hence, training an *x*-vector extractor on a limited amount of children's speech will result in sub-optimal performance. Data augmentation techniques offer a solution to these challenges. Data augmentation involves applying various transformations to the original training data to create new synthetic data samples. These synthetic samples are then used to augment the original data-set, thereby enhancing diversity of the captured acoustic attributes, increasing the amount of training data and improving the trained model's generalization capabilities. Taking cognizance of these facts, in-domain and out-of-domain data augmentation was performed to enhance the reliability and robustness of the developed children's ASV system.

### 2.1 Out-of-Domain Data Augmentation

Out-of-domain data augmentation refers to increasing the amount of training data by blending adults' data with children's speech. Since the acoustic attributes of adults' speech is in stark contrast to those of children, various modifications are applied to adults' speech so that the augmented data 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 outlined in Fig. 1. This augmentation technique involved using a limited quantity of original adults' speech. As noted earlier, we've used a variety of ways to adequately alter the acoustic characteristics of adults' speech. These are briefly addressed in the following:

In the first method, voice conversion (VC) was applied to the adults' speech using a cycle-consistent generative adversarial network (C-GAN) [11]. To train the C-GAN, about 10 minutes of speech samples from both adult and child speakers were employed.

As seen throughout the hearing tests, VC makes adult speech utterances sound remarkably similar to kid speech. As a result, the problems with acoustic mismatch are much reduced when the voice-converted data is pooled.

The second method was prosody modification applied to adult speech prior to augmentation. It is commonly known that children's speech has a higher pitch and a slower speaking tempo [15, 24]. As a result, the length of the speech data from the adult speakers was raised by 1.4 while the pitch was enhanced by a factor of 1.35. These scaling variables were chosen based on past studies that were published on children's speech recognition [27]. The method described in [22] was utilised to accomplish prosody modification (PM). To perform time-scale modification, the technique of audio stretching was applied, leveraging the methodology of fuzzy classification of spectral bins (FCSB) [4]. Again, pooling data that has been prosody-modified helps keep the acoustic mismatch under control.

Compared to adult speakers, formant frequencies are greater in the case of children [15, 24]. As a result, the formant frequencies (FM) of adult speech samples were scaled-up by a factor of 0.08 in the third approach. The aforementioned scaling factor was taken from previous publication [14]. Similar to VC and PM, pooling the data of formant modified adults' speech increases the training data while substantially reducing acoustic mismatch.

All the modified versions of adults' data were then pooled into training along with the original adults' data. A more reliable estimate of the model parameters was achieved as a result of increasing the training data volume. Furthermore, altering the acoustic characteristics makes sure that the established ASV system does not become biased towards speakers who are adults.

### 2.2 In-Domain Data Augmentation

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 speed perturbation technique was employed. The in-domain data augmentation technique is also pictorially represented in Fig. 1. Speed perturbation is one of the most well-known techniques for data augmentation reported in the scientific literature. In this technique the speaking-rate or speed is modified while preserving the linguistic content of the speech data. For this, the default three-way speed perturbation Kaldi pipeline is utilized. The speed of each of the utterances from children is modified simultaneously by a factor of 1.1 and 0.9, respectively. The speed perturbed data is then mixed with the unperturbed children's speech before learning the *x*-vector-based speaker representation.

### 2.3 Proposed Data Augmentation

The authors' in this paper propose a combination of the out-of-domain data augmentation as well as in-domain data augmentation as discussed in the previous subsections individually. All the modified versions of adults' and children's data are pooled into training along with original children's and adults' speech. Consequently, the proposed

data augmentation strategy addresses the challenges of acoustic variability posed by intra-speaker and inter-speaker variability, limited amount of training data, and potential adversarial attacks. The proposed data augmentation technique is pictorially summarized in Fig. 1. 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 Motivation for Exploring the Role of Feature Concatenation in Children ASV

As mentioned in the previous section, in the case of children, there is a considerable amount of relevant spectral information in the higher-frequency region. Children's speech data are represented by a spectrogram in the bottom panel of Fig. 3, which exhibits substantial power even between 4 and 8 kHz. Moreover, the spectrogram of children's speech fairly clearly illustrates the earlier literary works' assertion that the formant frequencies are higher in the case of child speakers [8, 13]. For a comparative study, Fig. 3 also includes the spectrograms for the speech data from adult male (top panel) and adult female (middle panel).

Mel-scale warping is influenced by the findings of psycho-acoustics. It is based on the premise that human perception of pitch is linear up to 1000 Hz and then becomes nonlinear for higher-frequencies [6]. 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. The down-sampling of spectral information in the high-frequency band is a snag when dealing with children's speech [8, 24]. The quest for the preservation of higher-frequency contents in children's speech led us towards the exploration of another front-end acoustic feature, namely the Inverse-Mel-Frequency Cepstral Coefficient.

The IMFCC features are extracted by projecting the power spectra onto inverse-Mel-weighted filter-banks. The inverse-Mel-filter-bank is realized simply by flipping around the Mel-weighted filter-banks about the middle point of the frequency axis. The configuration of inverse-Mel-filter-bank is depicted in the bottom panel of Fig. 2. The set up of this filter-bank is such that the high-frequency region's spectral information is better resolved and thus the IMFCC features are supposed to possess the acoustic attributes disregarded by the MFCC features. It is worth highlighting that due to the inherent nature of inverse-Mel-filter-bank, the spectral information in the lower frequency range of the children's speech will be down-sampled. Therefore, we have conceived the idea of concatenating the MFCC and IMFCC feature vectors in order to effectively preserve both the low as well as high-frequency components. The block diagram outlining the extraction process of the concatenated MFCC and IMFCC features is shown in Fig. 4.

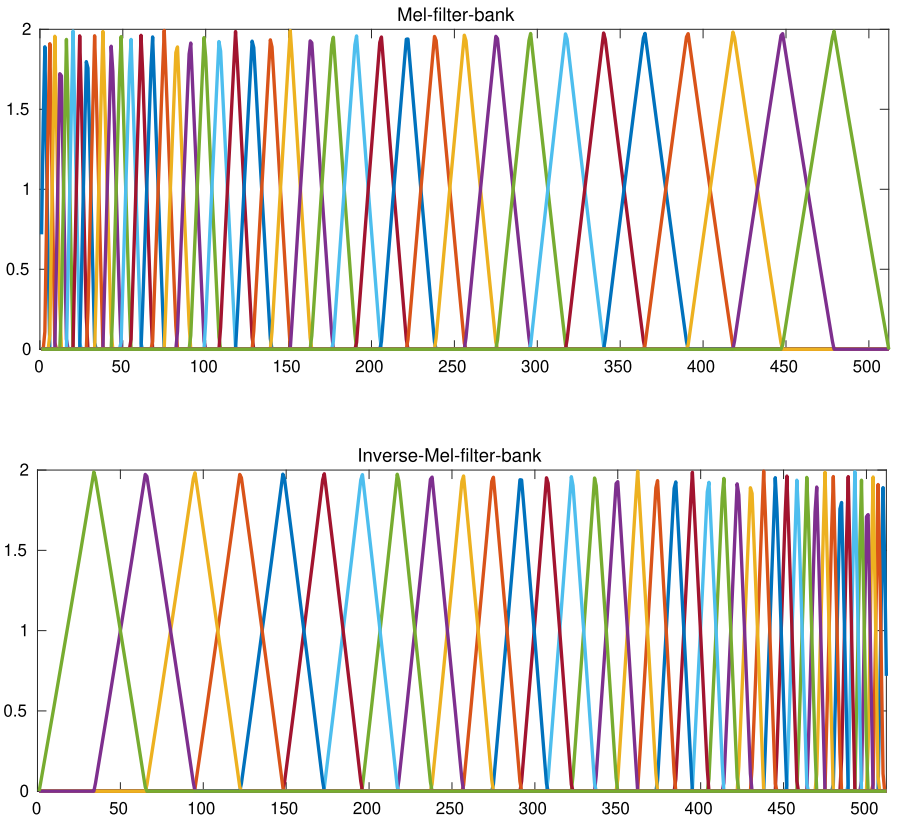


**Fig. 2** The configuration of a Mel- and Inverse-Mel filter-banks

### 3.1 Canonical Correlation Analysis (CCA)

Canonical Correlation Analysis (CCA) is a statistical technique used to analyze the relationship between two sets of variables. The primary goal of CCA is to identify and maximize the correlations between linear combinations of variables in each set. This is particularly useful when dealing with multiple variables in two data-sets and trying to understand the underlying relationships between them [16]. For a more comprehensive understanding of CCA, the appendix towards the end of this paper summarizes the procedure to compute CCA.

In order to substantiate the effect of feature concatenation, the CCA was carried out. We have computed the canonical correlation among MFCC and IMFCC features and the same is presented in Fig. 5. The CCA plot shows that the MFCC and IMFCC feature vectors are highly uncorrelated or less correlated for most of the coefficients except the starting few coefficients. Therefore the frame-level concatenation of MFCC and IMFCC features leads to capturing a wider range of acoustic attributes. The inherently complementary configuration of filter-banks employed in the extraction of MFCC and IMFCC features are the main force behind this development. Thus, the CCA plot of

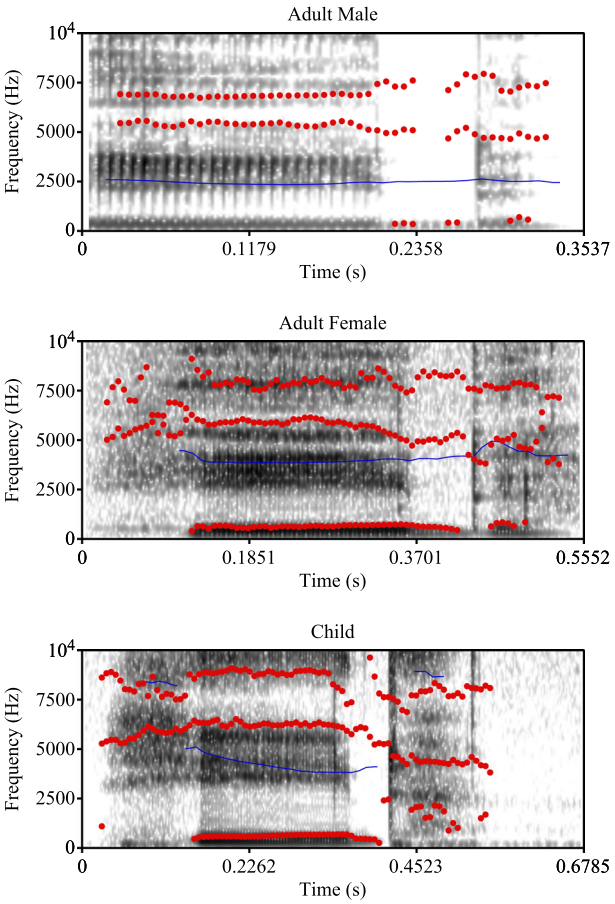


**Fig. 3** Spectrograms corresponding to speech data from adult male (top panel), adult female (middle panel) and child (bottom panel) speaking the word *HEED*. The red speckles are the contours denoting the variation in formant frequencies, while the blue line denotes the pitch frequency variations (Color figure online)

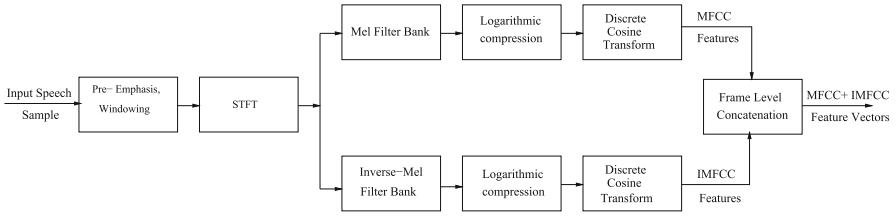


**Fig. 4** Block diagram outlining the process of extracting concatenated MFCC and IMFCC features

MFCC and IMFCC features upholds the complementary characteristic of IMFCC with respect to MFCC which assists their feature fusion model in representing a broader range of acoustic information in children's speech.

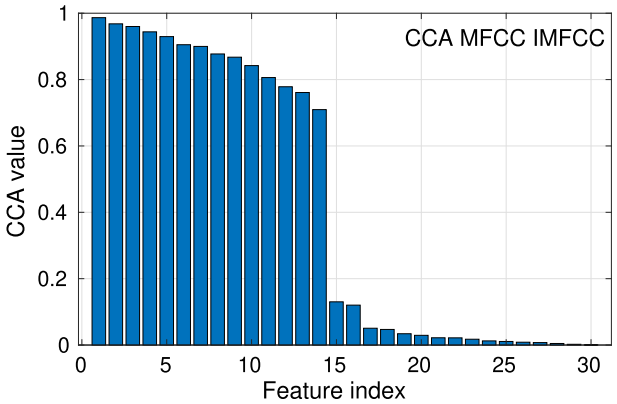


**Fig. 5** Canonical correlation analysis of MFCC and IMFCC feature concatenation explored in the paper

## 4 Experimental Evaluations

In this section, the relative effectiveness of MFCC, concatenated MFCC and IMFCC features are explored and the experimentally verified results are presented.

### 4.1 Employed Speech Corpora

Three different English speech corpora were employed for the development and evaluation of speaker verification system for children. These are CSLU kids corpus [31], CMU kids corpus [7], WSJCAM0 adults' speech corpus [23] and PF-STAR kids corpus [3]. The details of each data-set are tabulated in Table 1 and enumerated in the following:

i. *CSLU kids corpus*: This data set consists of spontaneous and prompted speech comprising of 100 hours of data having 73, 100 utterances from 1, 100 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.
ii. *CMU kids corpus*: This data-set comprises of 9.1 hour of data having 5, 180 utterances from 76 children. The child speakers are in the age group of 6 to 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 seconds. Therefore, evaluation on this set represents the short-utterance case.
iii. *WSJCAM0 corpus*: This adults' speech data-set is used for out-of-domain data augmentation. This corpus consists of 15.5 h of data with 7, 852 utterances, 132, 778 words from 92 adult speakers (male and female) having age exceeding 18 years. The sampling rate is 16 kHz.
iv. *PF-STAR kids corpus*: It consists of 8.3 h of data containing utterances from 121 child speakers aged between 4 and 14 years. To ensure consistency with the rest of our dataset, we down-sampled this data to 16 kHz from its original 44.1 kHz rate.

This down-sampled dataset serves as our test data for longer utterances. It includes 6,664 genuine trials and 995,420 impostor trials. On average, the duration of each data segment in this corpus is 30 s. Consequently, the evaluation on this set aims to highlight the contrast and demonstrate the severity of the problem when dealing with short utterances.

### 4.2 Experimental Set-Up

The Kaldi toolkit was used to create the entire ASV system configuration and perform all the experiments [20]. As already stated earlier, two front-end acoustic features, namely the MFCC and IMFCC are used to represent the speech signal. Both these features were extracted using the Kaldi toolkit. Speech data were passed through a first-order high pass filter, having pre-emphasis factor of 0.97. To bring stationarity in the nature of speech signals, the speech signal is examined separately in short time frames of 25 ms with an overlapping of 10 ms. A 30-channel Mel-filter-bank was utilized for projecting the power spectrum into Mel-scale, followed by the computation of the 30-dimensional MFCC features. While, for the computation of the IMFCC features, a 30-channel inverse-Mel filter-bank was employed for warping the power spectra to inverse-Mel-scale, before computing the 30-dimensional IMFCC 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 dataset 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-minute 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) [4];
iii. ADULT-FM: This data-set was generated by up-scaling the formant frequencies (FM) of ADULT speech by a factor of 0.08.

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

*Description of In-domain data augmentation:* The in-domain training set used for developing the children's ASV system was derived from children's speech data-set called the CSLU kids corpus. The default Kaldi pipeline for three-way speed perturbation [20] was utilized, wherein variations in speaking rate were introduced.

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

| Data used training | Duration of data(in Hrs) | Number of Speakers | Data used for testing | Duration of data(in Hrs) | Number of Speakers |
|---|---|---|---|---|---|
| CHILD (CSLU kids corpus) | 100 | 1100 | **CHILD (CMU kids corpus)** | **9.1** | **76** |
| CHILD + CHILD-SP | 300 | 3300 | | | |
| ADULT (WSJCAM0 corpus) | 15.5 | 92 | | | |
| ADULT-VC | 15.5 | 92 | | | |
| ADULT-PM | 31.5 | 184 | | | |
| ADULT-FM | 16 | 92 | | | |
| CHILD + ADULT + ADULT-VC + ADULT-PM + ADULT-FM | 178.5 | 1560 | | | |
| **CHILD+CHILD-SP+ADULT+ADULT-VC + ADULT-PM + ADULT-FM** | **378.5** | **3760** | | | |

Bold values indicate overall data used for training and testing purposes with their corresponding duration of data and number of speakers
The out-of-domain data augmentation scheme includes adult voice conversion (ADULT-VC), adult formant modification (ADULT-FM), adult prosody modification (ADULT-PM). The in-domain data augmentation scheme includes children's speech speed perturbation (CHILD-SP)

Specifically, each utterance of the CSLU kids corpus underwent simultaneous modifications in speed, accomplished through factors of 1.1 and 0.9, respectively. The speed perturbed children's speech is referred as CHILD-SP in this study. As a consequence of in-domain data augmentation, a total of 300 hours of speech data from child speakers (CHILD + CHILD-SP) is available for training purpose.

For the extraction of highly discriminative speaker representations, a deep neural network was utilised. These fixed dimensional speaker-embeddings called as $x$-vectors were extracted from a time-delay neural network (TDNN) architecture [33], comprising of 7 hidden layers and trained for 6 epochs. The training of network parameters was conducted utilizing the stochastic natural gradient descent algorithm [21, 33]. 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:

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

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).

### 4.3 Experimental Results

The foregoing sections of this article have dwelt in detail about the training of the ASV system on speech data comprising a large amount of original as well as perturbed children's speech along with an adequate amount of original and modified adults' speech database. To keep a track on the performance of the aforementioned ASV system when subjected to short utterances of children's test data-set, an investigative study was undertaken.

The first set of experiments were carried out to gauge the effectiveness the explored/proposed data augmentation techniques on the performance of the ASV system. The corresponding experimental results in terms of EER and minDCF are displayed in Table 2. As evident from the table, the performance evaluation metrics undergo successive improvement with the application of subsequent explored data augmentation techniques. For instance, with the application of out-of-domain data augmentation techniques the system records a relative improvement of 33.57% in EER with respect to the system trained on the child data-set alone. Next, when the in-domain data augmentation technique is put into action, the relative improvement in EER is 37.9%. Finally, as mentioned earlier, when the ASV system is trained using

**Table 2** EER and minDCF values for the short-utterances of children's speech test set demonstrating the effectiveness of out-of-domain, in-domain as well as the proposed data augmentation technique

| Type of Data augmentation | Data used for training | Evaluation Metrics EER (%) | minDCF |
|---|---|---|---|
| No data augmentation | CHILD | 21.95 | 0.9975 |
| Out-of-domain | CHILD + ADULT + ADULT-FM | 14.58 | 0.9233 |
| Data augmentation | + ADULT-PM + ADULT-VC | | |
| In-domain | CHILD + CHILD-SP | 13.63 | 0.9031 |
| Data augmentation | | | |
| In-domain + Out-of-Domain | CHILD+CHILD-SP+ADULT+ADULT-FM | **12.31** | **0.8464** |
| Data augmentation(**PROPOSED**) | + ADULT-PM + ADULT-VC | | |

Bold values indicate better performances achieved using the proposed approach in the paper
The out-of-domain data augmentation scheme includes adult voice conversion (ADULT-VC), adult formant modification (ADULT-FM), adult prosody modification (ADULT-PM). The in-domain data augmentation scheme includes children's speech speed perturbation (CHILD-SP)

**Table 3** Values of EER and the corresponding relative improvement in EER at each step of successive data augmentation techniques implemented to the employed ASV system

| Dataset | EER (%) | Relative improvement (%) |
|---|---|---|
| CHILD | 21.95 | – |
| CHILD + ADULT-FM | 19.78 | 9.88 |
| CHILD + ADULT-VC | 17.34 | 21.00 |
| CHILD + ADULT-PM | 16.30 | 25.74 |
| CHILD + ADULT-FM | 14.58 | 33.57 |
| + ADULT-PM + ADULT-VC | | |
| CHILD + CHILD-SP | 13.63 | 37.90 |
| **PROPOSED** | **12.31** | **43.91** |

Bold values indicate better performances achieved using the proposed approach in the paper

both the out-of-domain data augmentation technique and the in-domain data augmentation technique, which is the proposed data augmentation approach used in this paper, a staggering relative improvement of 43.91% with respect to the baseline system trained solely on child data-set is achieved. Consequently, the EER for the employed children's ASV system comes down to a measly 12.31%, which talks volumes about the effectiveness of the proposed data augmentation strategy. For more insight and to get a better feel of the efficacy of each of the consecutive data augmentation techniques, Table 3 lists the relative improvement in EER at each step of data augmentation.

It was mentioned earlier that the performance of a speaker verification system for children deteriorates when there is a reduction in the duration of the speech utterances during testing. The authors' take this opportune moment to demonstrate the detrimental effect of short-utterances on the performance of an ASV system. Table 4 compares the performance of an ASV system trained on the unperturbed child speech data, but tested on two different duration of children test speech samples: short-utterances and long

**Table 4** Equal error rate and minimum DCF values with respect to an *x*-vector-based ASV system trained on CHILD database and when evaluated using long and short utterances of children's speech test set

| Duration of test-set | EER (%) | minDCF |
|---|---|---|
| Long utterances | 6.38 | 0.7228 |
| Short Utterances | 21.95 | 0.9975 |

**Table 5** EER and minDCF values for the short-utterance-based ASV system trained on the data-set obtained using the proposed data augmentation technique demonstrating the effectiveness of feature concatenation

| Acoustic features | Evaluation metric | |
|---|---|---|
| | EER (%) | minDCF |
| MFCC | 12.31 | 0.8464 |
| MFCC + IMFCC | **11.30** | **0.8351** |

Bold values indicate better performances achieved using the proposed approach in the paper

**Table 6** 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 | 12.31 | 0.8464 |
| | 6–7 | 14.63 | 0.9657 |
| | 8–9 | 12.12 | 0.8481 |
| MFCC+IMFCC | Full test set | **11.30** | 0.8351 |
| | 6–7 | **13.61** | 0.9221 |
| | 8–9 | **10.98** | 0.8273 |

Bold values indicate better performances achieved using the proposed approach in the paper
This study was performed on *x*-vector-based ASV system trained on a mix of children's speech, adults' speech along with the modified versions of adults' and children's speech

utterances. As can be seen from Table 4, the performance of the *x*-vector-based system drops from 6.38% to 21.95%, for the baseline system trained solely on child data-set. This illustrates the significant difficulty presented by using short test utterances in ASV performance.

The next round of experiments were carried out to assess the effectiveness of the proposed frame-level concatenation of the two front-end acoustic features in the light of the employed short-utterance-based children's ASV system. The result of the evaluation metrics (EER and minDCF) obtained when MFCC features are concatenated with the IMFCC features for each frame of speech signal are shown in Table 5. It is to be kept in mind that the proposed data augmentation technique has been implemented prior to training the ASV system. The EER and minDCF values obtained when MFCC features alone are used to train the ASV system are also enlisted for comparison. Apparently, an absolute improvement of 1.01% is attained by the frame-level concatenation of MFCC and IMFCC features.

For an exhaustive analysis of the proposed strategy, the effect on the performance of the ASV system was monitored when subjected to an age-wise as well as gender-wise split-up of children's speech test set. For evaluating the effect of age variation, the

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

| Features | Child gender | EER (%) | minDCF |
|---|---|---|---|
| MFCC | Full test set | 12.31 | 0.8464 |
| | Female | 14.77 | 0.9113 |
| | Male | 9.229 | 0.7495 |
| MFCC+IMFCC | Full test set | **11.30** | 0.8351 |
| | Female | **13.75** | 0.8982 |
| | Male | **8.625** | 0.7311 |

Bold values indicate better performances achieved using the proposed approach in the paper

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

evaluation metric results are noted for the complete test set, as well as with split-up of the test-set in two subgroups on the basis of age. The corresponding values for EER and minDCF for this study are exhibited in Table 6. One should be mindful of the fact that the proposed data augmentation approach has been exercised prior to training the *x*-vector extractor. Going by the results of Table 6, it is quite evident that an ASV system shows a degraded results for children in the lower age bracket as compared to children in the higher age bracket or for the matter compared to children in the complete test-set. This can be attributed to the fact that younger children owing to their shorter vocal-track length have higher pitch frequency and formant frequencies. Also evident from the Table 6 is that the ASV system when trained exclusively on the MFCC features produce somewhat poorer results as those down-sample the higher-frequency contents of children's speech. On the contrary, the ASV system trained on the concatenated acoustic features yields superior results and this development can be attributed to the underlying fact that the feature fusion of MFCC and IMFCC takes into consideration the spectral information in the lower- as well as higher-frequency regions. Apart from the age-wise grouping of test set comprising children's short utterances, the effect of gender-wise grouping on the performance of the employed ASV system was also analyzed. The corresponding values for EER and minDCF for this study are given in Table 7. As noticeable from the table that the ASV system performance drops when subjected to female speech test-set in contrast to the male children or as opposed to children in the complete test-set. This deterioration is due to the higher formant and pitch frequencies of female child in comparison to male child. Table 7 again echos the superior performance of the ASV system trained on the concatenated MFCC and IMFCC features as against the system trained solely on MFCC features.

Moving on from the qualitative analysis towards the quantitative analysis of the effect of the proposed feature concatenation on the employed children's ASV system. The EER for the full test set registers a relative improvement of 8.20% when MFCC features are concatenated with the IMFCC features, pictorially represented by the first bar in Fig. 6. When the speech test-set is split on the grounds of age variation, a relative reduction in EER for the age bracket of 6–7 years is calculated as 6.97% when the

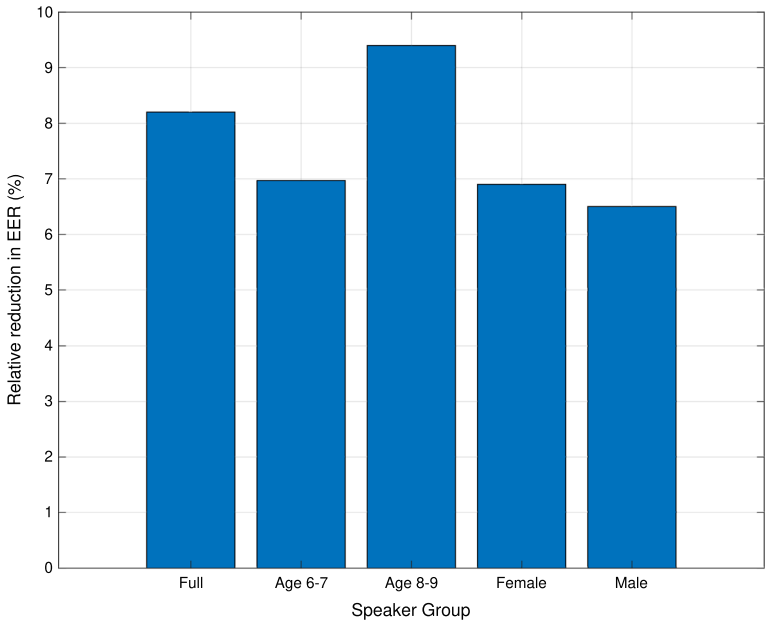


**Fig. 6** Bar graph representation of the relative reduction in EER (%) for various speaker groups (in terms of age and gender), corresponding to the ASV system trained on the concatenated MFCC and IMFCC features as compared to an ASV system trained on the MFCC features alone

proposed feature concatenation is put into action. This is depicted by the second bar in the Fig. 6. The corresponding relative improvement in EER for the age bracket of 8–9 years is 9.40% portrayed by the third bar in Fig. 6. When the speech test-set is split on the grounds of gender, a relative reduction in EER for the girl child is calculated as 6.90% upon frame-level concatenation of MFCC with the IMFCC features, depicted by the fourth bar of the Fig. 6. Finally when the employed ASV system is subjected to short-utterances from the speech test set of the male child, it results in a relative reduction of 6.50% in EER. The same has been pictorially represented by the fifth bar of the Fig. 6.

## 5 Conclusion

Through the work in this paper, the authors' have examined the challenges surrounding the task of building a children's speaker verification system and their potential applications. Firstly, it was evident that the traditional speaker verification techniques designed for adult speakers are not directly applicable to children due to their physiological and morphological differences. The development of a robust and reliable children ASV system requires abundance of domain-specific data-set. This requirement was met by the proposed data augmentation strategy employed in this paper. Incorporating both in-domain and out-of-domain data augmentations in the proposed data augmentation approach, the amount of training data was increased, the diversity of the captured acoustic attributes was widened which also led to an improvement in the trained model's generalization capabilities, while keeping the acoustic mismatch

in check. A relative improvement of 43.91% in equal error rate (EER) against the baseline system trained solely on the original child data-set authenticates the potency of the proposed data augmentation approach. Together with data augmentation, the effectiveness of frame-level concatenation of MFCC with the IMFCC features, is also analysed in this paper. The complementary nature of filter-banks employed in the extraction of IMFCC and MFCC features, helps in preserving spectral information in the higher-frequency range. The frame-level concatenation of the MFCC and IMFCC features results in a relative reduction of 8.20% for the complete test-set. Additionally, age- and gender-wise analyses were carried out to study the combined effect of data augmentation and feature concatenation on the performance of children's ASV system. The ASV system incorporating both the proposed data augmentation technique as well as feature concatenation culminates in an impressive overall relative improvement of 48.51% for equal error rate. 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 towards improved security, personalized experiences, and educational opportunities for children while ensuring their safety and well-being.

**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.

**Funding** Funding information is not applicable/No funding was received.

**Ethics 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.

## Appendix

### Procedure to Compute Canonical Correlation Analysis (CCA)

A brief explanation of the procedure to compute CCA is summarized in the following:

i. Formulation of Hypothesis: CCA starts with the formulation of a hypothesis that there exist canonical variables (linear combinations of the original variables) in each dataset that are highly correlated.
ii. Data Collection: Collect two sets of variables, typically denoted as X and Y. These sets may represent different measurements, features, or attributes.
iii. Standardization (Optional): Standardize the variables in both sets if necessary. This step is optional, but it can be useful to ensure that variables are on a similar scale.
iv. Construct Covariance Matrices: Create covariance matrices for both sets (Cov(X) and Cov(Y)).

v. Compute Cross-Covariance Matrices: Calculate the cross-covariance matrix between X and Y (Cov(X, Y)).
vi. Solve Generalized Eigenvalue Problem: Solve the generalized eigenvalue problem derived from the covariance matrices. This involves finding the eigenvalues and corresponding eigenvectors of the matrix equation $Ax = \lambda Bx$, where A is the cross-covariance matrix, and B is the product of the inverse square root of the covariance matrices of X and Y.
vii. Canonical Variables: The canonical variables are the linear combinations of the original variables that maximize the correlation. These are obtained from the eigenvectors of the generalized eigenvalue problem.
viii. Canonical Correlations: The canonical correlations are the square roots of the eigenvalues obtained in the previous step. These represent the strength of the relationships between the canonical variables.
ix. Interpretation: Interpret the canonical variables and correlations to understand the underlying relationships between the two sets of variables. Higher canonical correlations indicate stronger relationships.

## References


1. S. Aziz, S. Shahnawazuddin, Effective preservation of higher-frequency contents in the context of short utterance based children's speaker verification system. Appl. Acoust. **209**, 109,420 (2023)
2. K. Badillo-Urquiola, D. Smriti, B. McNally, E. Golub, E. Bonsignore, P.J. Wisniewski, Stranger danger! social media app features co-designed with children to keep them safe online. in *Proceedings of the 18th ACM International Conference on Interaction Design and Children*, pp. 394–406 (2019)
3. A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, M. Wong, The PF_STAR children's speech corpus. in *Proceedings of INTERSPEECH*, pp. 2761–2764 (2005)
4. E.P. Damskägg, V. Välimäki, Audio time stretching using fuzzy classification of spectral bins. Appl. Sci. **7**(12), 1293 (2017)
5. S. D'Arcy, M. Russell, A comparison of human and computer recognition accuracy for children's speech. in *Ninth European Conference on Speech Communication and Technology* (2005)
6. 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
7. M. Eskenazi, J. Mostow, D. Graff, The CMU Kids Corpus LDC97S63. https://catalog.ldc.upenn.edu/LDC97S63 (1997)
8. M. Gerosa, D. Giuliani, S. Narayanan, A. Potamianos, A review of ASR technologies for children's speech. in *Proceedings of Workshop on Child, Computer and Interaction*, pp. 7:1–7:8 (2009)
9. R.M. Hanifa, K. Isa, S. Mohamad, A review on speaker recognition: technology and challenges. Comput. Electr. Eng. **90**, 107005 (2021)
10. A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, Plda based speaker recognition on short utterances. in *Proceedings of The Speaker and Language Recognition Workshop: Odyssey 2012*, pp. 28–33. International Speech Communication Association (2012)
11. T. Kaneko, H. Kameoka, Parallel-data-free voice conversion using cycle-consistent adversarial networks. arXiv preprint arXiv:1711.11293 (2017)
12. 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 (2020)
13. 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)

14. 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)
15. 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)
16. M. Observations, Multivariate observations, gaf seber, ed (1984)
17. V. Peddinti, D. Povey, S. Khudanpur, A time delay neural network architecture for efficient modeling of long temporal contexts. in *Proceedings of INTERSPEECH* (2015)
18. 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)
19. A. Poddar, M. Sahidullah, G. Saha, Quality measures for speaker verification with short utterances. Digital Signal Process. **88**, 66–79 (2019) https://doi.org/10.1016/j.dsp.2019.01.023
20. D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, K. Vesely, The Kaldi Speech recognition toolkit. in *Proceedings of ASRU* (2011)
21. D. Povey, X. Zhang, S. Khudanpur, Parallel training of deep neural networks with natural gradient and parameter averaging. in *Proceedings of ICLR* (2015)
22. S.R.M. Prasanna, D. Govind, K.S. Rao, B. Yegnanarayana, Fast prosody modification using instants of significant excitation. In *Proceedings of International Conference on Speech Prosody* (2010)
23. T. Robinson, J. Fransen, D. Pye, J. Foote, S. Renals, WSJCAM0: a British English speech corpus for large vocabulary continuous speech recognition. in *ICASSP 1995–1995 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 81–84 (1995)
24. M. Russell, S. D'Arcy, Challenges for computer recognition of children's speech. in *Proceedings of Speech and Language Technologies in Education (SLaTE)* (2007)
25. M. Russell, S. D'Arcy, L. Qun, The effects of bandwidth reduction on human and computer recognition of children's speech. IEEE Signal Process. Lett. **14**(12), 1044–1046 (2007)
26. S. Safavi, M. Russell, P. Jancovic, Automatic speaker, age-group and gender identification from children's speech. Comput. Speech Language, **50** (2018)
27. S. Shahnawazuddin, N. Adiga, H.K. Kathania, B.T. Sai, Creating speaker independent asr system through prosody modification based data augmentation. Pattern Recognit. Lett. **131**, 213–218 (2020). https://doi.org/10.1016/j.patrec.2019.12.019
28. 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 Process. **93**, 34–42 (2019)
29. 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)
30. S. Shahnawazuddin, W. Ahmad, N. Adiga, A. Kumar, Children's speaker verification in low and zero resource conditions. Digital Signal Process. **116**, 103115 (2021)
31. K. Shobaki, J.P. Hosom, R. Cole, Cslu: Kids' Speech Version 1.1. Linguistic Data Consortium (2007)
32. S. Snyder, D. Garcia-Romero, D. Povey, S. Khudanpur, Deep neural network embeddings for text-independent speaker verification. in *Proceedings of INTERSPEECH*, pp. 999–1003 (2017)
33. D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, X-vectors: robust DNN embeddings for speaker recognition. in *ICASSP 2018–2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333 (2018)
34. M. Tsujikawa, T. Nishikawa, T. Matsui, I-vector-based speaker identification with extremely short utterances for both training and testing. in *2017 IEEE 6th Global Conference on Consumer Electronics (GCCE)*, pp. 1–4. IEEE (2017)
35. G. Yeung, A. Alwan, On the difficulties of automatic speech recognition for kindergarten-aged children. Interspeech 2018 (2018)