Circuits, Systems, and Signal Processing (2023) 42:6228–6252
https://doi.org/10.1007/s00034-023-02399-y

Feature Fusion and Ablation Analysis in Gender Identification of Preschool Children from Spontaneous Speech

Kodali Radha[1] · **Mohan Bansal**[1]

Received: 11 March 2022 / Revised: 4 May 2023 / Accepted: 5 May 2023 /
Published online: 20 May 2023

Abstract

The children below 6 years of age are called preliterate who use speech as one of their primary forms of communication. Fundamental frequency or pitch is a characteristic that is used to classify gender, but young children have reasonably similar pitch due to their immature vocal tract which varies from 215 to 390 Hz for both genders. Most studies for gender identification have utilized pitch and mel frequency cepstral coefficients (MFCC), because of their ability to capture the efficacy of signals. However, the performance of pitch and MFCC on noisy speech signals are poor, and as a result, they fail to accurately detect gender characteristics. Considering this limitation, the proposed work investigates the novel fusion and ablation experimentation of mel frequency cepstral coefficients (MFCC) and gamma-tone frequency cepstral coefficients (GFCC). To enhance the accuracy of a robust text-independent children gender identification model, the cepstral features are combined with the tonal descriptors (pitch and harmonic ratio). The most contributing front-end features were selected by fusion and ablation analysis and distributed to a bagged tree classifier ensemble. To manage the memory requirements, redundant features are trimmed using principle component analysis (PCA). The hyper-parameter optimization is accomplished using the grid search technique to further increase frame-level accuracy. This study is likely to be a forerunner in the field of children's speech recognition, which has been revealed to be a reliable and accurate method of gender identification.

Keywords Gender identification · Pitch · Harmonic ratio · MFCC · GFCC · MF-GFCC · LOFO · Ensemble of bagged trees

✉ Mohan Bansal
mohan.bansal@vitap.ac.in

Kodali Radha
kodaliradha.20PHD7093@vitap.ac.in

[1] School of Electronics Engineering, VIT-AP University, Amaravati, Andhra Pradesh 522237, India

1 Introduction

Children's automatic speech recognition (ASR) systems have lagged behind adult ASR in terms of reliability. The precise challenges and strategies for evaluating child ASR were yet to be substantially investigated. Recent research from the robotics community reveals that ASR for kindergarten speech is exceptionally hard [31], regardless of the fact that voice-based pedagogical and diagnostic tools may benefit this age group the most. The review of grade and gender-specific ASR systems was performed, with an emphasis on kindergarten-aged children (5–6 years old) [16]. Gender classification of children is one of the most significant procedures in speech processing. The paralinguistic effects can be identified from an acoustic speech signal and used to infer the speaker's identity, gender, age, personality traits, accent, and emotional states [22–24, 27]. For example, in some situations like a telephone conversation, it is difficult to identify the gender of children [5]. The difficulty of distinguishing a child's gender and the presence of multiple significant speech cues between boys and girls have garnered attention in gender classification research. For the OGI Kid's corpus, a complete and in-depth investigation of the automatic speaker, age group, and gender identification from children's speech was presented [27]. However, it may not have been able to account for all possible factors that could affect gender identification from children's speech, such as cultural background, accent, or language. Another research [26] presents different combinations of features used to evaluate the efficiency using machine-learning techniques, which allows for flexibility in gender identification. Moreover, additional research is needed to determine the most effective features for gender identification in children's speech. A technique for developing gender and age-based automatic models of 174 children (between 6 and 11 years old) was detailed in another article [18]. The aforementioned study illustrates that gender identification results have greatly improved compared to recent findings by various other researchers, but it's challenging to correctly classify individuals who are extremely similar in age. Therefore, our research aims to fill this gap and can accurately classify the gender in children below 6 years of age.

The proposed research addresses these concerns by utilizing feature fusion and ablation, as well as hyper-parameter tuning. The fusion and ablation of mel frequency cepstral coefficients (MFCC) and gamma-tone frequency cepstral coefficients (GFCC) feature vectors are concatenated with pitch and harmonic ratio. Then modeled using an ensemble classifier for gender detection in an attempt to significantly improve the representation of features. The process of fusing MFCC and GFCC is known as "feature fusion", whereas "feature ablation" is the process of removing undesirable features to investigate individual features influencing the performance of the model.

The organization of the article is as follows. Section 2 discusses related work, Sect. 3 articulates the speech data sets used in this article, and Sect. 4 describes the proposed methodology, which involves feature extraction using tonal and cepstral variations, and classification using ensemble learning, Sect. 5 presents the experimental setup, and Sect. 6 presents the experimental results. Finally, Sect. 7 concludes the proposed empirical work briefly with future scope.

2 Related Work

There has been considerable work on gender classification of adults' speech and achieved high accuracy of 99% [1, 4, 8], but quite little work done on children's speech. Extracting the para-linguistic information of children like age, gender, dialect, and emotions from the speech are challenging. Some reasons may include the fact that small vocal tracts of children, high levels of spectral variability, high pitch and resonant frequencies, and lack of children's speech data [32]. The study of gender identity from the pitch (f_0) and formant frequencies ($F_1 + F_2 + F_3$) of children's speech aged between 4–16 years was attempted in [19] with an accuracy of 74%. Later on, many researchers reported the effectiveness of pitch in distinguishing gender in children. A typical approach found that the fundamental frequency f_0 is high in spontaneous speech [17] under noisy environments which is another challenge in the present model. In [6], three voice sources: cepstral peak prominence, harmonic-to-noise ratio and spectral harmonic magnitudes were extracted for gender detection and compared with MFCC for 5 age groups of children between 8–17 in conjunction with support vector machine (SVM) and Gaussian mixture model (GMM) achieved an accuracy ranging from 61–91%. In the article [26], the MFCCs, linear predictive cepstral coefficients (LPCC), formants, pitch, shimmer and jitter features are extracted from reading speech of children aged 6–11 and gained an accuracy of 84.79% with random forest (RF) classifier. Recently, [2] employed CatBoost machine to choose features from MFCC and spectral subband centroids (SSC) in age and gender classification of children aged 7 to 14, achieving an accuracy of 86.23% with SVM classifier. Although the classification accuracy of male and female increases with the age of the children because of the maturity in vocal folds, however, considering the decrease in age groups as suggested in this model is relatively difficult, this inspired us to reliably classify the gender of preschool children from their speech.

Finally, the main contribution of this article is the classification of gender in children below 6 years of age (preliterate children) from their spontaneous speech, which is still an elusive and very little explored task. In addition, to complete this task, a novel feature fusion and ablation analysis of MFCC and GFCC were used to investigate the most promising characteristics in differentiating male and female classes of children, as well as pitch and harmonic ratio in conjunction with the model ensemble.

3 Speech Data Sets

The database used in this study is recorded English speech at a preschool in the U.K with 11 kindergarten school children consisting of 5 females and 6 males with a mean (μ) age of 4.9 years and standard deviation (σ) of 4 months which was developed by child speech recognition in human–robot interaction [15]. The dataset contains 670 recordings, out of which 280 from females and 390 from males. Children's speech is collected during interaction with robots holding three microphones (NAO, PORT and STUDIO) placed at different natural locations in the school and includes noisy environments like fan noise, bird noise, door closing, other children shouting from other classrooms, etc.

The dataset consists of three different categories of speech utterances: single-word sentences, fixed sentences, and spontaneous speech. The single word utterances were numbers from 1 to 10, fixed utterances were 5 short sentences (e.g., "*the boy climbed out of the window*"), and spontaneous speech consists of long sentences which were collected through story retelling from a picture book

All the recordings are in the format of "*.wav*" with a sampling frequency of 44.1 kHz, resulting in 16 files per microphone (48 in total) per child. The spontaneous speech was transcribed and cut into 222 sentences of various lengths ($\mu = 7.8$ words per utterance, $\sigma = 2.6$). However, the selection of this database in the proposed approach was centered on the possibility of identifying the gender of children from spontaneous speech rather than short utterances, which is useful for market applications of ASR.

4 Proposed Methodology

High-performance gender identification systems rely heavily on data pre-processing, robust feature extraction, and, finally, classification. After thorough experimentation of selecting optimum features from fusion and ablation of MFCC and GFCC compared against pitch and harmonic ratio, the proposed work is the only one to apply GFCC in the identification of child's gender. The best features are used to train an ensemble of bagged tree classifiers, and gender class is predicted via majority voting.

4.1 Spontaneous Speech in Children

One of the most challenging aspects in children speech recognition is obtaining better accuracy when spoken responses are relatively unscripted or spontaneous and ungrammatical (e.g., "*The girl putted the box on the table*"). Preschool children speech (4–6 years) is remarkably different from higher grade children (above 7 years) and identification performance is only 45.5% whereas higher grade children achieved 75% [7]. Each recording of spontaneous speech used in this study is around 2–3 min long and contains natural noise settings from several microphones.

4.2 Speech Pre-Processing

Pre-processing is mostly used to remove undesirable noise from an audio source. Speech is a non-stationary signal with a low frequency (upto 4 kHz), but is expected to be stationary for a short period of time, i.e., between 20 and 30 ms. Pre-emphasis [12] is a straight-forward signal processing technique that boosts the amplitude of a high-frequency speech signal to make it stronger than noise, which indeed improves the signal-to-noise ratio (SNR).

The output of the pre-emphasis network is represented as E that boosts the high-frequency components of an audio signal. It is typically applied to speech signals before further processing, such as compression or recognition. Equation (1) shows how the output of the pre-emphasis network at nth sample is calculated. It is the difference

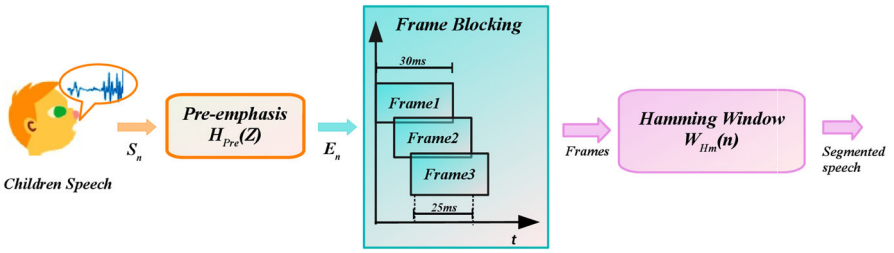


Fig. 1 Segmentation of children speech (S_n) using pre-emphasis filter ($H_{\text{Pre}}(Z)$) and frame duration of 10 ms with a hamming windowing ($W_{\text{Hm}}(n)$) of 30 ms duration

between the current sample of the input speech signal (S_n), and the previous sample (S_{n-1}) multiplied by coefficient α, as depicted in Eq. (1).

$$E_n = S_n - \alpha S_{n-1} \tag{1}$$

The finite impulse response (FIR) of the high-pass pre-emphasis filter ($H_{\text{Pre}}(Z)$) is obtained from Eq. (1) by taking the Z-transform of both sides and is given in Eq. (2).

$$H_{\text{Pre}}(Z) = \frac{E(Z)}{S(Z)} = 1 - \alpha Z^{-1} \tag{2}$$

where 'Z' represents the Z-transform variable, and 'α' is a pre-emphasis coefficient in the range of $0 < \alpha \leq 1$.

The process of splitting the pre-emphasized audio signal into a sequence of frames is referred to as framing and windowing [29]. Therefore, in this paper 30 ms hamming window $W_{\text{Hm}}(n)$ with a 25 ms overlap length to avoid signal loss, are used for segmentation as shown in Fig. 1. The hamming window is a tapering function that smooths out the start and end discontinuities of each sample using Eq. (3).

$$W_{\text{Hm}}(n) = 0.54 - 0.46 \cos\left(\frac{2 \cdot \pi \cdot n}{N}\right), \quad 0 \leq n \leq N \tag{3}$$

where 'N' denotes the number of samples.

4.3 Proposed Tonal Descriptors

This section provides a detailed explanation of the pitch, harmonic ratio, and zero crossing rate as efficient tonal descriptors that can be derived from children's speech.

4.3.1 Pitch (Fundamental Frequency)

The study of tonal characteristics in speech, such as pitch and harmonic ratio, is essential for understanding the prosody of children's speech. Pitch, specifically the fundamental frequency, is considered to be a crucial indicator of gender discrimination

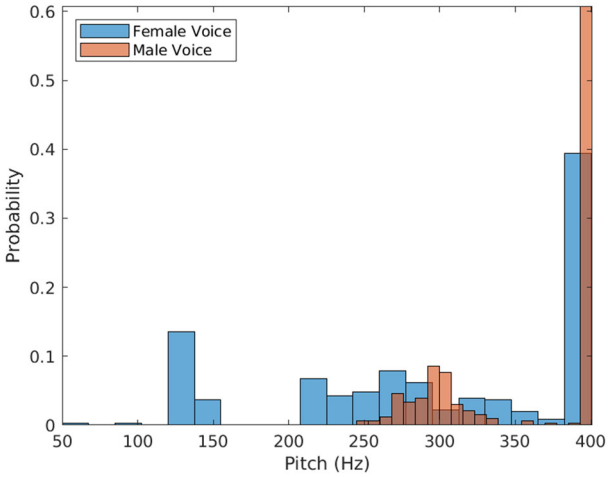


Fig. 2 Histogram representation of pitch for male and female children

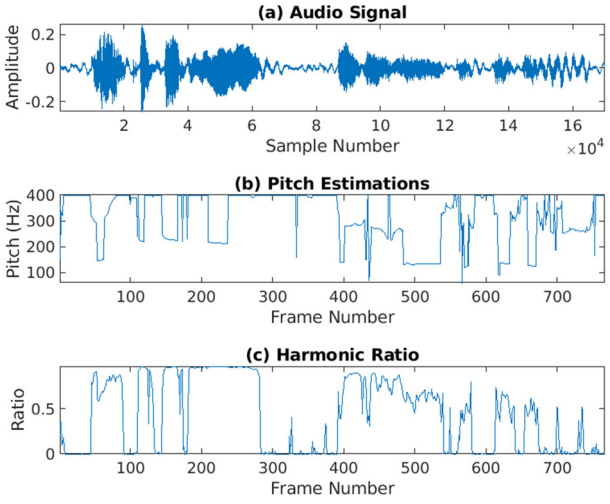


Fig. 3 Comparison of pitch and harmonic ratio: **a** represents an example audio signal **b** represents the pitch or fundamental frequency of the given signal **c** represents the harmonic ratio of the given signal

and identification in children's speech. Previous research has found that children have a higher pitch due to their immature vocal tracts, with ranges varying from 210 to 415 Hz for children between the ages of 3 and 7 [11, 16]. This range is typically lower for female children, which is between 212 and 375 Hz ($\mu = 290 \pm \sigma = 26$) and slightly higher for male children, which is between 245 and 390 Hz ($\mu = 315 \pm \sigma = 25$). However, the similarity in pitch ranges between the two genders makes gender classification challenging as shown in Fig. 2. Therefore, this study has explored the use of another prosodic trait, the harmonic ratio, in combination with pitch to aid in detecting children's gender as shown in Fig. 3.

Pitch or fundamental frequency (f_0) as shown in Eq. (4), is determined for a given input frame of 30 ms hamming window with an overlap of 25 ms [30].

$$f_0 = \frac{1}{2L}\sqrt{\frac{\varphi}{\rho}}, \tag{4}$$

where L effective vocal fold length, φ longitudinal stress, and ρ tissue density.

4.3.2 Harmonic Ratio

In 1977, Stephanie Seneff introduced a method for detecting the harmonic ratio (HR) by analyzing the distance between harmonics in a specific region of the spectrum, specifically below 1100 Hz [28]. This technique is used to derive the fundamental frequency (f_0) of a sound by analyzing the harmonics created by the vibration of the vocal cords and the flow of acoustical air from the vocal tract. The harmonic ratio ($\hbar_r$) is a measure of the number of frequency components in the power spectrum of a sound. For each frame of the input signal, the HR is estimated using Eq. (5), and the resulting feature vectors are then normalized using an auto-correlation function (ACF).

$$\hbar_r = \frac{\sum_{n=1}^{N} S_n S_{n-i}}{\sqrt{\sum_{n=1}^{N} S_n^2 \sum_{n=0}^{N} S_{n-i}^2}}, \quad \text{for} \quad 1 \leq i \leq \top \tag{5}$$

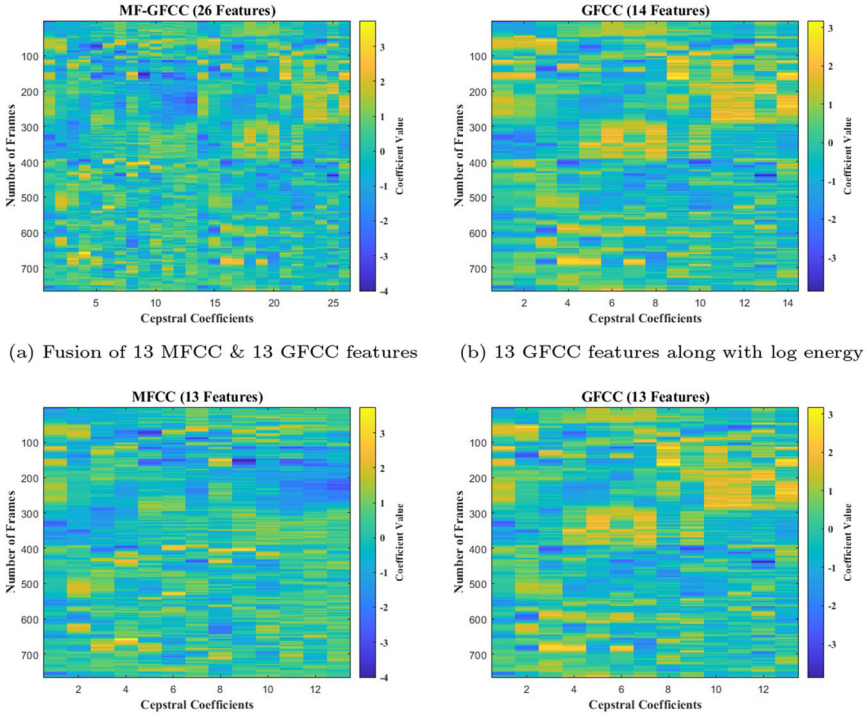
where, S_n denotes speech signal at nth sample with N frame length, i is the delay of ACF and $\top$ is the maximum delay referring to the minimal f_0.

4.3.3 Zero Crossing Rate

The speech waveform can be divided into three regions: voiced, unvoiced, and silence. The voiced speech region is considered to be the most informative, as it contains valuable information on the pitch and harmonic ratio, which can be used to distinguish it from silence and unvoiced speech [3]. To separate silence and speech regions, a short-time power threshold is calculated, and the zero crossing rate (ZCR) is determined using Eq. (6) to distinguish between voiced and unvoiced speech [29]. It is possible to identify the voiced speech regions by combining the power threshold and ZCR for each frame.

$$\text{ZCR} = \frac{1}{N}\sum_{n=1}^{N}\left[1 - \text{sgn}\{S_n S_{n-1}\}\right], \tag{6}$$

where $\text{sgn}\{\cdot\} = \begin{cases} 1 & S_n S_{n-1} \geq 0 \\ 0 & \text{Otherwise} \end{cases}$ S_n is the nth sample value of the speech signal and frame period N.



(a) Fusion of 13 MFCC & 13 GFCC features (b) 13 GFCC features along with log energy

(c) 13 MFCC features by ablation of log energy (d) 13 GFCC features by ablation of log energy

Fig. 4 Feature fusion and feature ablation of cepstral descriptors: **a** represents the 26 static feature vectors of both MFCC and GFCC extracted from all frames, **b** represents the 13 static GFCC features along with log energy (2nd coefficient), **c** represents, the ablation of log energy resulting 13 static MFCC features, **d** represents, the ablation of log energy resulting 13 static GFCC features

4.4 Cepstral Descriptors

To investigate the relevant cepstral features in speech for children's gender identification, the MFCCs and GFCCs are used to compare with the pitch and harmonic features. In general, the classification performance is measured by the quality of the set of feature vectors. As a result, irrelevant features may affect the gender identification model to be less accurate. Obtaining a specific set of linguistic features is of high priority in machine learning to achieve better classification performance. Furthermore, many researchers reported that MFCC features provide the best possible results in gender identification, but that MFCC features are less accurate if the children's speech coincides with noise. In particular, introducing GFCC feature engineering in children's gender identification under natural environmental settings, and an effective combination of MFCC fused with GFCC features and GFCC features alone with an appropriate learning mechanism yields state-of-the-art results. The fusion and ablation of front-end features are shown in Fig. 4. This illustrates the fusion and ablation of front-end features. Figure 4a displays the fusion of MFCC and GFCC with 13 static features, where a clear distinction between the coefficients can be seen. Figure 4b

shows GFCC with log energy located at the 2nd cepstral coefficient. Figure 4c, d depicts the ablation of log energy, showing only 13 static features. It is notable that in comparison to Fig. 4b, d, the 2nd cepstral coefficient was removed from the 14 GFCCs, resulting in the clear visibility of 13 static GFCCs in Fig. 4d.

4.4.1 MFCC Feature Vectors

There is rich literature on extracting features by using MFCCs due to the ability to simplify the speech amplitude spectrum in a cosine form on a nonlinear mel scale [14, 29]. To capture the dynamic features of non-stationary speech, Δ and $\Delta\Delta$ coefficients along with static MFCCs ($\hat{\mathbb{C}}_i$) were estimated using Eqs. (7) and (8).

$$\hat{\mathbb{C}}_i = \sum_{j=1}^{k} (\log \hat{S}_j) \cos\left[i \left(j - \frac{1}{2} \right) \frac{\pi}{k} \right], \tag{7}$$

where $\hat{\mathbb{C}}_i$ is the ith cepstral coefficients of static MFCC, $\log \hat{S}_j$ denotes the log filter bank amplitudes of speech and 'k' is number of filter bank channels. The following equation is to calculate Δ(differential) feature coefficients,

$$\Delta_{i,t} = \frac{\sum_{i=1}^{k} i[\hat{\mathbb{C}}_{i,t} - \hat{\mathbb{C}}_{i,t-1}]}{2\sum_{i=1}^{k} i^2} \tag{8}$$

Here $\Delta_{i,t}$ and $\hat{\mathbb{C}}_{i,t}$ are the ith delta coefficient and cepstral coefficients of MFCC for frame 't' respectively. $\Delta\Delta_{i,t}$ (acceleration) feature coefficients can be directly calculated from Δs using Eq. (9).

$$\Delta\Delta_{i,t} = \Delta_{i,t} - \Delta_{i,t-1} \tag{9}$$

4.4.2 GFCC Feature Vectors

GFCC features, on the other hand, are extracted because MFCCs are susceptible to noise. MFCCs alone cannot provide better accuracy because children's speech datasets were gathered in noisy environments. The gamma-tone frequency cepstral coefficients (GFCC) are a group of gamma-tone filter banks that are used to mimic the auditory features of a human ear. The gamma-tone filter bank output can be used to create a cochleagram, which is a time-frequency representation of the signal and it is comprised of a number of overlapping bandpass filters that are specified in the temporal domain by its impulse response at particular points all along the cochlea. The gamma tone distribution function and sinusoidal function form an impulse response in the time domain as shown in Eq. (10).

$$\gamma(t) = \hat{A} \cdot t^{\eta-1} \cdot \exp(-2\Pi W t) \cos(2\Pi f_c t + \Phi) \tag{10}$$

Here $\hat{A}$ controls the output gain, W represents the bandwidth, η represents the order of the gamma tone filter that defines the gradient of the edges and which is usually set to an order less than 4, f_c(Hz) is the central frequency and Φ(radians) is the phase shift often set to 0.

Similar to the triangular filter banks of MFCC, gamma-tone filter banks of GFCC are the discrete cosine transform of nonlinear cubic root of the equivalent rectangular bandwidth (ERB) scale [20]. The static features of GFCC can be computed from Eq. (11).

$$\hat{\gamma}_i = \sqrt{\frac{2}{\eta}} \sum_{j=0}^{\eta} \frac{1}{3} (\log \hat{S}_j) \cos \left[i \left(j - \frac{1}{2} \right) \frac{\pi}{\eta} \right] \tag{11}$$

where $\hat{\gamma}_i$ is the ith cepstral coefficients of static GFCC, $1/3 \log(\hat{S}_j)$ denotes the cubic root of log filter bank amplitudes of speech and η is a number of filter bank channels. The following Eqs. (12) and (13) are used to calculate $\eth_{i,t}$(Differential) feature coefficients and $\eth\eth_{i,t}$(Acceleration) feature coefficients from static GFCC.

$$\eth_{i,t} = \hat{\gamma}_{i,t} - \hat{\gamma}_{i,t-1} \tag{12}$$

$$\eth\eth_{i,t} = \eth_{i,t} - \eth_{i,t-1} \tag{13}$$

4.5 Ensemble Learning Classifier

Ensemble learning is a generic machine learning meta-approach that aims to improve predictive performance by aggregating predictions from individual models. Compared to individual decision tree classifiers, the ensemble methods can significantly improve the classification performance in children's speech.

The ensemble model's basic principle is that a cluster of weak learners joins together to become strong learners, boosting the model's accuracy. The predictions made by the ensemble members are then aggregated using simple statistics, such as voting as depicted in Algorithm 1. Although there are nearly an unlimited number of ways, to choose the best parameters of the bagged tree [21] ensemble classifier shows high accuracy for all the tested cases, and thus, it is chosen for the purpose of child gender classification in this article.

Assume $X_i \in \{x_1, x_2, \ldots, x_n\}$ to be the original feature set with $n \times q$ matrix, where x_n is the n^{th} sample with q features and $Y_j \in \{y_1, y_2\}$ is the class labels in classification of gender, where y_1 is the female class label and y_2 is the male class label. The bootstrap samples from randomly selecting with replacement are denoted by $\Theta_b^* \in \{\theta_1^*, \theta_2^*, \ldots, \theta_B^*\}$, where B as a number of learner models.

Consequently, bootstrap samples are modeled using B decision tree classifier and denoted as $C_{DT}^*(X_b) = \text{Info}(\Theta_b^*)$ which is the information gain [25] of each bootstrap data. The model estimator of an ensemble of bagged tree classifier is given as $\Theta_{\text{bag}}^* = \underset{y \in Y}{\text{argmax}} \left\{ \sum_{b=1}^{B} C_{DT}^*(X_b) = Y_j \right\}$ that chooses the output class which receives the highest votes as final classification.

Algorithm 1 Bagged tree classifier for child gender classification

Input: Child speech dataset [Train 75%, Test 25%]
Use 5-fold cross validation.
$[X_i, Y_j]$ Training data X_i with labels $Y_j \epsilon \{ y_1, y_2\}$
Result: Female child & male child classification

for i= 1:n **do**
$\Theta_b^* = [X_i^*, Y_j^*]$ ▷ Bootstrap sampling of $[X_i, Y_j]$
$C_{DT}^*(X_b) = \text{Info}(\Theta_b^*)$ ▷ Information gain of bootstrap data
end for
for j= 1:2 **do**
$\Theta_{bag}^* = \underset{y \in Y}{\text{argmax}} \left\{ \sum_{b=1}^{B} C_{DT}^*(X_b) = Y_j \right\}$ ▷ Majority vote prediction
end for

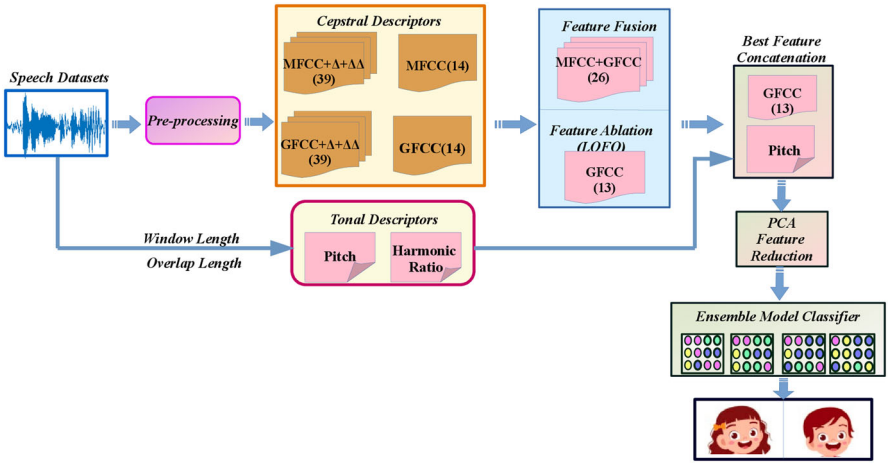


Fig. 5 Proposed methodology using feature fusion and ablation for gender identification in pre-school children

In this work, the ensemble of bagging tree classifier is implemented with the total number of splits are 62,623 and the number of learners are 30. Classification of male and female in children can be inevitably identified using spontaneous speech recordings of subjects as shown in Fig. 5.

5 Experimental Setup

The experimental setup for developing the gender identification model using the proposed feature fusion and feature ablation of MFCC and GFCC using ensemble learning is presented in this section. In-depth investigations were conducted in four distinct ways to evaluate the performance of the suggested child gender identification model.

5.1 Feature Fusion and Ablation Analysis

In this research, both cepstral and tonal descriptors are extracted from children's spontaneous speech. Spectral descriptors are mainly static and dynamic MFCCs and GFCCs, as well as fused MF-GFCC features, whilst the pitch and harmonic ratio are tonal descriptors.

5.1.1 Feature Fusion Analysis

Fusion of MFCC and GFCC features is a technique used to combine the strengths of both feature sets and improve the performance of gender identification systems for children's speech. MFCCs are a commonly used feature in speech processing systems, as they capture the spectral characteristics of speech in a compact form. They are based on a nonlinear frequency scale, which makes them efficient to compute and easy to interpret. However, they may not be optimal for certain types of speech, such as children's speech, where the characteristics of the voice may be different from adult speech. Children's voices are known to have different pitch, formants, and prosodic patterns than adult speech, which can make them harder to identify. GFCCs, on the other hand, are designed to better handle non-stationary signals, such as those found in children's speech. They are based on a nonlinear cubic root of ERB frequency scale, which allows them to capture the harmonic structure of speech. This is important for capturing the unique characteristics of children's voices, such as their pitch, formants, and prosodic patterns.

By concatenating MFCC and GFCC features, a gender identification system can take advantage of the strengths of both feature sets and improve its performance on children's speech. The MFCCs can provide a good representation of the spectral characteristics of speech, while the GFCCs can capture the harmonic structure of speech. The combined feature set can be more robust and accurate than using a single feature set alone.

In practice, the concatenation is done by appending the GFCCs to the MFCCs, forming a longer feature vector that will be used as an input to the gender identification model. This concatenated feature vector can be used to train and test the model. Once the model is trained, it can be used to perform gender identification on children's speech with a high degree of accuracy and robustness. Basic acoustic features (BAF) were extracted from the children's speech initially, which included static MFCC along with Δ and $\Delta\Delta$ (39) feature vectors and static GFCC along with Δ and $\Delta\Delta$ (39) feature vectors. For the comparison against BAF, static MFCC and static GFCC (26) feature vectors were concatenated in the feature fusion (FF) setup.

In addition, it's also worth noting that gender identification from children's speech is a complex task as it relies on both acoustic and linguistic features, therefore, fusing different features from multiple sources can help the system to be more robust and generalize better.

5.1.2 Feature Ablation Analysis

Feature ablation, particularly leave-one-feature-out (LOFO) [9], is important in children's gender identification from speech because it allows for the evaluation of the contribution of each individual feature to the overall performance of the gender identification system. LOFO is a method of feature ablation where each feature is removed one at a time and the performance of the gender identification system is evaluated with the remaining features. This allows for the identification of which features are most important for accurate gender identification and which features can be removed without significantly impacting performance. In the case of children's gender identification from speech, different features such as pitch, formants, harmonic ratio, acoustic, and prosodic features may have different importance. By removing each feature one at a time, it is possible to understand which features are most important in children's speech, and which ones may not be necessary.

This analysis can be useful in identifying redundant or less important features, which can then be removed from the feature space, resulting in a smaller and more efficient feature space. This can also help in understanding the contribution of each feature to the final decision of the model. Additionally, it can aid in identifying which feature type (MFCC or GFCC) is more important for the specific task, which in turn can help to improve the generalization ability of the system. The ablation experiment that was part of the proposed investigation is represented by the FA1 and FA2 experiments. Thirteen static features of MFCC or GFCC along with log energy are the outcomes of the ablation of dynamic features (Δ & $\Delta\Delta$) of MFCC or GFCC in FA1 whereas FA2 represents the feature selection of 13 static features of MFCC or GFCC by ablation of log energy.

Overall, feature ablation is a powerful tool for understanding the underlying mechanisms of gender identification in children's speech and for improving the performance of gender identification systems.

5.2 Transform Best FA2 Features Vectors using Principle Component Analysis (PCA)

An excessive intake of concatenated features will decelerate the computational speed of the model. As a result, using dimensionality reduction algorithms like PCA to remove redundant features is crucial. To improve the performance of the proposed model in terms of frame-level accuracy, PCA is enabled. The PCA reduces 14 features to 9 features that account for 90% of variability. The training time and prediction speed before and after enabling PCA were recorded.

5.3 Hyper-Parameter Tuning

The hyper-parameter optimization or tuning in machine learning is the process of choosing a group of optimal hyper-parameters for a learning algorithm. In this paper, grid search is used to fine-tune the ensemble model's parameters search with learning

rate ranges between 0.001–1, number of predictors to be 10 features, and learners to be 10–500.

5.4 Comparison of Proposed State-Of-the-Art Features with Baseline Results

To show the effectiveness of the proposed model, a comparison with baseline models was performed. The extracted features of UK children's speech datasets are incorporated and retrained by other baseline models like artificial neural network (ANN), deep neural network (DNN), random forest (RF), support vector machine (SVM), and multi-layer perceptron (MLP).

6 Experimental Results

This section reveals the results of all the experiments discussed in Sect. 5.

6.1 Performance Analysis using Feature Fusion and Ablation Study

In this section, to evaluate the classification performance of gender identification in children's speech and to verify the choice of modeling, different combinations of cepstral and tonal descriptors were used as shown in Table 1. The experiments are conducted in such a way that the combination of cepstral features are concatenated with pitch and harmonic ratio separately and predicted the accuracy of the child gender identification model in terms of files (File_{Acc}) and frames ($\text{Frame}_{\text{Acc}}$).

- *Gender Identification (GI) Model using Pitch* (f_0) *Concatenation:*
  The basic acoustic characteristics (BAF), which are composed of 39 MFCC feature vectors and 39 GFCC feature vectors, were taken into consideration in the initial experiment. 39 MFCCs were concatenated with a pitch that yields a File_{Acc} of 89.85% and $\text{Frame}_{\text{Acc}}$ of 85.7%. Similarly, the combination of pitch and 39 GFCC features yielded a File_{Acc} of 92.75% and $\text{Frame}_{\text{Acc}}$ of 86.47%. Since GFCCs are inherently more noise-robust than MFCCs, their adoption enhances the performance of the GI model. Further, the investigation into feature fusion (FF) of 13 static MFCC and 13 static GFCC (MF-GFCC) continued and concatenated with a pitch, which results in File_{Acc} of 94.2% and $\text{Frame}_{\text{Acc}}$ of 89.3%. When comparing the FF study with the BAF study of GFCC, it has been shown that there is a relative improvement (RI) in the accuracy of 1.52% at the file level and 3.27% at the frame level. A feature ablation (FA1) investigation is being performed in the subsequent iteration by ablating the dynamic features ($13\Delta + 13\Delta\Delta$) from the BAF of MFCCs and GFCCs. This produces 13 static and the co-existing log energy features concatenated with the pitch showing that in the MFCC, File_{Acc} rapidly drops to 85.5% and $\text{Frame}_{\text{Acc}}$ to 83.34%. Additionally, the empirical use of the GFCC (13 static features + log energy) concatenated with a pitch for gender prediction demonstrates the higher accuracy at the file and frame level reporting, at 98.55 and 90.11%, respectively, as demonstrated in Fig. 6. The final experimental setup aimed to evaluate the impact of each feature on the model's performance. To

Table 1 Experimental results of gender identification system using fusion and ablation of cepstral features when combined with pitch (f_0) and harmonic ratio (HR), in terms of file accuracy ($\text{File}_{\text{Acc}}\%$) and frame accuracy ($\text{Frame}_{\text{Acc}}\%$) respectively

	Cepstral descriptors (no. of features)	Tonal descriptors			
		Pitch (f_0)		Harmonic ratio (HR)	
		$\text{File}_{\text{Acc}}(\%)$	$\text{Frame}_{\text{Acc}}$ (%)	File_{Acc} (%)	$\text{Frame}_{\text{Acc}}$ (%)
Basic acoustic features (BAF)	MFCC (39 features)	89.85	85.7	89.85	85.22
	GFCC (39 features)	92.75	86.47	92.75	86.26
Feature fusion (FF)	MFCC + GFCC (13 + 13 Static features)	94.2	89.3	92.75	89.16
Feature ablation (FA1)	MFCC (13 Static features + Log Energy)	85.5	83.34	85.50	83.25
	GFCC (13 Static features + Log Energy)	98.55	90.11	98.55	89.7
Feature ablation (FA2) (leave one feature out)	MFCC (13 Static features)	85.5	83.68	85.5	83.25
	GFCC (13 Static features)	100	90.13	98.55	89.7

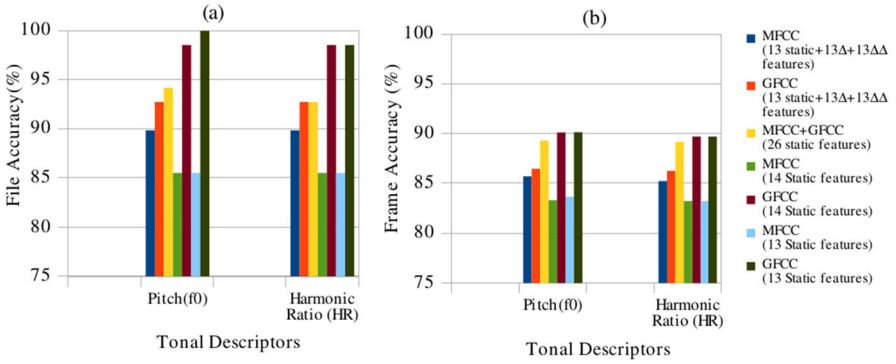


Fig. 6 Performance analysis: **a** Effect of fusion and ablation of cepstral features on file accuracy (%) with pitch (f_0) and harmonic ratio (HR), separately. **b** Effect of fusion and ablation of cepstral features on frame accuracy (%) with pitch (f_0) and harmonic ratio (HR), separately

achieve this, a leave-one-feature-out (LOFO) ablation study, also known as FA2, was conducted. The study involved removing one feature at a time from the 14 features of MFCC and GFCC. These 14 features included 13 static cepstral features and log energy. The model was then retrained to assess its performance. The results of the LOFO study showed that in the case of MFCCs concatenated with pitch, removing any MFCC feature or log energy did not affect the model's accuracy. However, in the GFCC setup of FA2, excluding the log energy feature resulted in an improvement in accuracy. Thus, it was concluded that all cepstral features of GFCC contributed to the model's performance, except for log energy. Log energy added noise or redundancy to the feature set, which negatively affected the model's ability. The results demonstrated a significant enhancement in the model's performance when the log energy feature was removed from the GFCCs. As a result, when log energy was excluded from GFCCs and combined with pitch, the accuracy per file reached 100% and the accuracy per frame was 90.13%, as indicated in Table 1.

- *Gender Identification (GI) Model using Harmonic Ratio (HR) Concatenation:* As it was evident in the above section, GFCC performed effectively when combined with pitch. The subsequent harmonics of fundamental pitch, known as the harmonic ratio, are taken into consideration and evaluated for gender prediction in order to advance the research. Initially, the BAF, which is composed of 39 MFCC feature vectors and 39 GFCC feature vectors, was taken into consideration. 39 MFCCs were concatenated with a HR that yields almost similar performance as pitch which reported to a File_{Acc} of 89.85% and $\text{Frame}_{\text{Acc}}$ of 85.22%. Similarly, the combination of HR and 39 GFCC features yielded a File_{Acc} of 92.75% and $\text{Frame}_{\text{Acc}}$ of 86.26%. This BAF research revealed that the cepstral features combined with pitch performance are equally similar to the cepstral features concatenated with HR performance. Further research into the feature fusion (FF) of 13 static MFCC and 13 static GFCC (MF-GFCC) was conducted and combined with the HR which results in File_{Acc} of 92.75% and $\text{Frame}_{\text{Acc}}$ of 89.16%. When comparing the FF study with the BAF study of GFCC, it has been shown that there

Table 2 Confusion matrix of gender identification model of feature ablation studies FA1 & FA2 in terms of $\text{File}_{\text{Acc}}(\%)$

Class	FA1: GFCC (14 features) + Pitch		FA2: GFCC (13 features) + Pitch	
	F	M	F	M
F	96	4	100	0
M	0	100	0	100

F female child, *M* male child

is no relative improvement (RI) at the file level, but a RI of 3.36% is observed at the frame level. A feature ablation (FA1) results in 13 static and log-energy features combined with HR have revealed that, in the MFCC, File_{Acc} and $\text{Frame}_{\text{Acc}}$ both swiftly decrease to 85.5 and 83.25%, respectively. Additionally, GFCC (13 static features + 1 log energy) concatenated with a HR for gender prediction demonstrates higher accuracy at the file and frame level reporting, at 98.55 and 89.7%, respectively. In the final stage of the experiment, the impact of harmonic ratio and LOFO (FA2) characteristics on the performance of the model was evaluated. The results showed that the addition of HR and LOFO did not affect the accuracy of the FA2 model, which was consistent with that of the previous stage, FA1, as demonstrated in Table 1. This finding suggests that the HR feature was already highly discriminative and played a crucial role in accurately classifying audio signals. Furthermore, when the log energy feature was removed from the cepstral features of MFCCs and GFCCs and concatenated with HR, there was no significant impact on the model's performance, which remained at 98.55% at the file level and 89.7% at the frame level.

The gender identification performance of GFCC + Pitch from FA1 and FA2 experiments are further examined in the confusion matrix of Table 2. In FA1, 4% of the female class is misclassified as male which results in 98% of overall classification accuracy, and leaving a Log Energy from 14 features of GFCC results in 100% accuracy of the proposed model.

6.2 Performance Analysis using Feature Transformation with PCA

As a part of reducing the computational complexity of the suggested model, PCA transformation of key contributing features of FA2 was followed. As demonstrated in the above experiment, GFCC (13 static features) + pitch (1 feature) achieves 100% accuracy (File_{Acc}) and 90.1% ($\text{Frame}_{\text{Acc}}$), but the space requirements in MATLAB are rather extremely high due to the high dimensionality of the feature space. However, the typical goal of using PCA is to mitigate the redundant feature dimensions from 14 to 9 while incorporating required components that explain 90% of the variability, resulting in a new set of 9 features known as principal feature components.

An important measure in the selection of a suitable feature subspace called "Explained Variance" is considered to be 90% which corresponds to those top 9 features. The explained variance per component in descending order is constructed with

Table 3 Performance analysis of ensemble model before and after PCA in terms of number of features, frame accuracy (%), prediction speed (observations/second), and training time (seconds)

Parameters	Before PCA	After PCA
No. of features	GFCC + Pitch (13 static + 1)	GFCC + Pitch (8 static + 1)
Accuracy/frame	90.13%	90.13%
Prediction speed	42,000 obs/s	45,000 obs/s
Training time	51.59 s	46.051 s

Table 4 List of hyper-parameters used in grid-search optimization in tuning the ensemble model

Parameter	Optimal configuration value
Optimizer	Grid search
Number of grid divisions	10
Number of iterations	100
Number of learners	210
Learning rate	0.02155
Maximum number of splits	5380
Observed minimum classification error	0.048

33.3%, 14.5%, 9.9%, 7.8%, 7.0%, 6.2%, 4.6%, 4.1%, 3.7% and are more than enough to describe the entire data set. And discarding the remaining variance of those not opted 5 features contains the least information and can be excluded. To inspect the application of PCA in the proposed model, prediction speed is increased and training time is reduced as shown in Table 3. However, it has been discovered that the same accuracy can be achieved even with reduced features.

6.3 Performance Analysis of Optimizable Ensemble Model

The relative importance of the proposed gender identification model, which employs the grid search algorithm of hyper-parameter tuning, is examined in this section. Tuning the hyper-parameters of an ensemble classifier, such as the learning rate, the maximum number of splits, and the number of learners, of a machine learning algorithm, influences the learning process of a model to work at an optimal level for performance improvement. As a result, various hyper-parameter tuning approaches, such as grid search, Bayesian search, and random search, are frequently used [10]. The Bayesian and random search methods are sluggish and unsophisticated. Hence, the grid search method was used to select the optimum hyper-parameters for the proposed model.

Using the grid search method, we develop a model for various hyper-parameter combinations, validate the model for each combination, and save the findings. The selection of hyper-parameters that yields the desired results out of all possible combinations is identified as the best parameter set for the proposed model. Using a grid

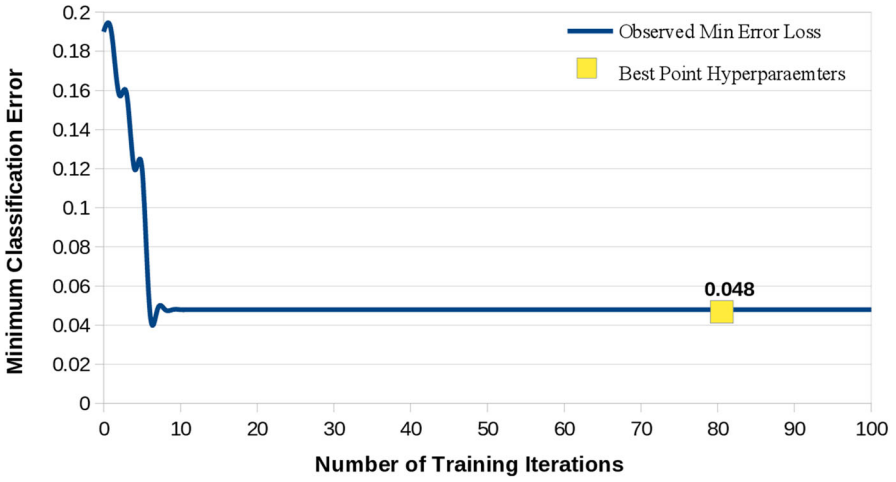


Fig. 7 Grid-search hyper-parameter optimization

search approach, the determined best set of hyper-parameters are shown in Table 4 with a minimum classification error [13] of 0.048. The yellow spot in Fig. 7 represents the iteration that corresponds to the hyper-parameters that provide the observed minimum classification error (MCE).

6.4 Comparative Analysis of Proposed System with Earlier Work

To evaluate the performance of the proposed system for gender identification in children, various front-end features used in previous studies were compared in Table 5 based on their classification accuracy. However, it's important to note that the results obtained by the proposed method cannot be directly compared with the baseline methods due to differences in datasets. The baseline models used in these studies were trained on datasets that are not publicly available, making direct comparison difficult. Instead, to assess the effectiveness of the proposed approach, the most contributing features (GFCC + Pitch) from UK children's speech datasets were fed to other baseline models, as presented in Table 6. While these models showed improvement over the compared method, with a relative improvement of 1.04% using ANN and 2.23% using DNN, as reported in [26], the proposed ensemble of bagged trees achieved even higher accuracy. The bagged trees model outperformed the baseline models and achieved the highest accuracy, highlighting its effectiveness for gender identification in children.

In nutshell, the ensemble model proposed in this study, which uses GFCC and pitch features, has shown promising results for identifying the gender of preschool children. Additionally, incorporating these features into various baseline models demonstrates the effectiveness of our proposed approach in improving the accuracy of gender classification.

Table 5 Comparative analysis with earlier children gender identification models

Authors	Front-end feature extraction	Corpus type (age)	Modeling approach	Classification results (accuracy)%
[19]	$f_0 + F_1 + F_2 + F_3$	7 non-diphthongal vowels of American English (4–16)	Statistical analyses-ANOVA simple correlations, and multiple regression analysis	74%
[6]	MFCC + CPP + HNR + f_0 and formant values	English children-read speech with vowel change (8–17)	SVM, GMM	61–91%
[33]	MFCC+ $f_0 + F_1 + F_2 + F_3$	Malaysian children (7–12)-vowels	MLP, HMM	99.81%
[27]	MFCC+$\Delta + \Delta\Delta$	English children (5–16)-spontaneous and read speech	GMM-UBM, GMM-SVM i-vectors with PLDA	79%
[18]	MFCC + f_0 + Spectral descriptors + POV + ZCR	Private datasets (6–11) spontaneous speech	ConvNets	71%
[26]	MFCCs, LPCCs, formants pitch, shimmer and jitter	English children-read speech (6–11)	Artificial neural network (ANN)	76.20%
			Deep neural network (DNN)	78.25%
			Random forest (RF)	84.21%
[2]	MFCC+SSC	English children-(7–14)	SVM	86.23%
Proposed work	Feature fusion and ablation of features	English children spontaneous speech (5–6)		
			Ensemble of bagged tree model	94.2%
	MFCC + GFCC + Pitch			98.55%
	GFCC (13 Static + 1 Log Energy) + Pitch			98.55%
	GFCC (13 Static + 1 Log Energy) + HR			100%
	GFCC (13 Static) + Pitch			

Table 6 Retraining baseline models with proposed state-of-the-art front-end features

Model used in baselines	Parameter setup	Baseline model accuracy on proposed features (%)	Relative improvement	Remarks
ANN [26]	Number of fully connected layers: 1	77%	1.04%	The author used 68 feature vectors in classification of children gender, but this work uses only 14 features to obtain the relative improvement of 1.04%
	First layer size: 25			
	Activation: ReLU			
	Iteration limit: 1000			
DNN [26]	Number of fully connected layers: 3	80.1%	2.23%	Relative improvement of 2.23% occurred when baseline model DNN is retrained with current datasets with GFCC+Pitch feature extraction
	3 layer size: 10			
	Activation: ReLU			
	Iteration limit: 1000			
RF [26]	Maximum number of splits: 4	70%	–	Observed high accuracy in baseline model of random forest
	Split criterion: Gini's diversity index			
	Surrogate decision splits: Off			
SVM [2]	Kernel function: Linear	68%	–	Observed high accuracy in baseline model of linear SVM
	Kernel scale: Automatic			
	Box constraint level: 1			
	Multi-class method: One-vs-one			

Table 6 continued

Model used in baselines	Parameter setup	Baseline model accuracy on proposed features (%)	Relative improvement	Remarks
MLP [33]	Number of fully connected layers: 1 First layer size: 25 Activation: Sigmoid Iteration limit: 1000 Regularization strength (Lambda): 0	81.1%	–	Since author used vowel based classification, the accuracy would be obviously high compared to utterance or frame level classification
Proposed model	Ensemble method: Bagging Learner type: Decision tree Maximum Number of splits: 62,623 Number of learners: 30	100%	–	Proposed model outperforms the existing work on UK children spontaneous speech datasets using GFCC + Pitch front end features

7 Conclusion and Future Work

In this paper, the noise-robust front-end features are extracted from the voiced samples of spontaneous speech datasets. Novel fusion and ablation studies of MFCC and GFCC that combine the effectiveness of tonal descriptors (pitch and harmonic ratio) were employed to detect the gender of children. The rigorous experimental finding revealed that the performance of the various combinations of features shows an overall accuracy of approximately 89.8–100% outperforms the previous approaches. Moreover, the GFCC and pitch were found to be suitable for classifying gender using an ensemble model of bagged trees. The performance of the model was evaluated by reducing the feature dimensions using PCA and furthermore focusing on memory requirements and training time. The ideal parameter set is determined via the grid search hyperparameter optimization, with a least miss-classification error of 0.048. The existing approaches for determining children's gender either demands clean speech datasets or require more feature extractors because of more spectral and temporal changes. It is important to say that the data is from 3 different microphones (NAO, PORT, and STUDIO) in a noisy environment with targeted children of below 6 years of age whose speech has higher spectral and acoustic variability. Despite multiple constraints, this innovative methodology is able to predict the gender of children with an accuracy of 100% outperforming existing methods. The obtained results motivate us to continue researching ensemble learning, which might be used as a universal model for classifying para-linguistic effects in large datasets of spontaneous speech in children.

Acknowledgements The authors would like to extend their gratitude to VIT-AP University for providing the essential resources necessary to conduct this research at the High-Performance Computing Laboratory.

Funding This research received no external funding.

Data Availability Statement The open access data that support the findings of this study is available from the ZENODO repository, "https://zenodo.org/record/200495#.Yit0zXpBxPZ". More details about the data are given in Sect. 3.

Declaration

Conflict of interest The authors declare no conflict of interest.

References

1. R.S. Alkhawaldeh, DGR: gender recognition of human speech using one-dimensional conventional neural network. Sci. Program. (2019)
2. A.A. Badr, A.K. Abdul-Hassan, CatBoost machine learning based feature selection for age and gender recognition in short speech utterances. Int. J. Intell. Eng. Syst. **14**(3), 150–159 (2021)
3. M. Bansal, P. Sircar, Parametric representation of voiced speech phoneme using multicomponent AM signal model, in *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)* (IEEE, 2018), pp. 128–133
4. M. Bansal, P. Sircar, Phoneme based model for gender identification and adult-child classification, in *13th International Conference on Signal Processing and Communication Systems (ICSPCS)* (IEEE, 2019), pp. 1–7

5. T. Bocklet, A. Maier, J.G. Bauer et al., Age and gender recognition for telephone applications based on GMM supervectors and support vector machines, in *ICASSP* (IEEE, 2008), pp. 1605–1608
6. G. Chen, X. Feng, Y.L. Shue et al., On using voice source measures in automatic gender classification of children's speech, in *Eleventh Annual Conference of the International Speech Communication Association* (2010)
7. T. Cincarek, I. Shindo, T. Toda et al., Development of preschool children subsystem for ASR and Q&A in a real-environment speech-oriented guidance task (2007)
8. F. Ertam, An effective gender recognition approach using voice data via deeper LSTM networks. Appl. Acoust. **156**, 351–358 (2019)
9. D. Feng, F. Chen, W. Xu, Efficient leave-one-out strategy for supervised feature selection. Tsinghua Sci. Technol. **18**(6), 629–635 (2013)
10. M. Feurer, F. Hutter, *Hyperparameter Optimization* (Springer, Cham, 2019), pp.3–33
11. E.J. Hunter, A comparison of a child's fundamental frequencies in structured elicited vocalizations versus unstructured natural vocalizations: a case study. Int. J. Pediatr. Otorhinolaryngol. **73**(4), 561–571 (2009)
12. R. Jahangir, T.Y. Wah, N.A. Memon et al., Text-independent speaker identification through feature fusion and deep neural network. IEEE Access **8**, 32,187-32,202 (2020)
13. B.H. Juang, W. Hou, C.H. Lee, Minimum classification error rate methods for speech recognition. IEEE Trans. Speech Audio Process. **5**(3), 257–265 (1997)
14. H.K. Kathania, S. Shahnawazuddin, N. Adiga et al., Role of prosodic features on children's speech recognition, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018), pp. 5519–5523. https://doi.org/10.1109/ICASSP.2018.8461668
15. J. Kennedy, S. Lemaignan, C. Montassier et al., Child speech recognition in human–robot interaction: evaluations and recommendations, in *Proceedings of the 2017 ACM/IEEE International Conference on Human–Robot Interaction* (2017), pp. 82–90
16. 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)
17. A. McAllister, S.K. Brandt, A comparison of recordings of sentences and spontaneous speech: perceptual and acoustic measures in preschool children's voices. J. Voice **26**(5), 673.e1-673.e5 (2012)
18. H. Pérez-Espinosa, H. Avila-George, J. Martínez-Miranda et al., Children age and gender classification based on speech using ConvNets. Res. Comput. Sci. **147**, 23–35 (2018)
19. T.L. Perry, R.N. Ohde, D.H. Ashmead, The acoustic bases for gender identification from children's voices. J. Acoust. Soc. Am. **109**(6), 2988–2998 (2001)
20. J. Qi, D. Wang, J. Xu et al., Bottleneck features based on gammatone frequency cepstral coefficients, in *Interspeech, International Speech Communication Association* (2013)
21. J.R. Quinlan et al., Bagging, boosting, and c4. 5, in *AAAI/IAAI*, vol. 1 (1996), pp. 725–730
22. K. Radha, M. Bansal, Audio augmentation for non-native children's speech recognition through discriminative learning. Entropy **24**(10), 1490 (2022)
23. K. Radha, M. Bansal, Closed-set automatic speaker identification using multi-scale recurrent networks in non-native children. Int. J. Inf. Technol. **15**(3), 1375–1385 (2023)
24. K. Radha, M. Bansal, S.M. Shabber, Accent classification of native and non-native children using harmonic pitch, in *2022 2nd International Conference on Artificial Intelligence and Signal Processing (AISP)* (IEEE, 2022), pp. 1–6
25. L.E. Raileanu, K. Stoffel, Theoretical comparison between the Gini index and information gain criteria. Ann. Math. Artif. Intell. **41**(1), 77–93 (2004)
26. P.B. Ramteke, A.A. Dixit, S. Supanekar et al., Gender identification from children's speech, in *2018 Eleventh International Conference on Contemporary Computing (IC3)* (IEEE, 2018), pp. 1–6
27. S. Safavi, M. Russell, P. Jančovič, Automatic speaker, age-group and gender identification from children's speech. Comput. Speech Lang. **50**, 141–156 (2018)
28. S. Seneff, Real-time harmonic pitch detector. IEEE Trans. Acoust. Speech Signal Process. **26**(4), 358–365 (1978). https://doi.org/10.1109/TASSP.1978.1163118
29. G. Sharma, K. Umapathy, S. Krishnan, Trends in audio signal feature extraction methods. Appl. Acoust. **158**(107), 020 (2020)
30. Y.L. Shue, M. Iseli, The role of voice source measures on automatic gender classification, in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. (IEEE, 2008), pp. 4493–4496
31. G. Yeung, A. Alwan, On the difficulties of automatic speech recognition for kindergarten-aged children. Interspeech (2018)

32. F. Yu, Z. Yao, X. Wang et al., The SLT 2021 children speech recognition challenge: Open datasets, rules and baselines, in *2021 IEEE Spoken Language Technology Workshop (SLT)* (IEEE, 2021), pp. 1117–1123
33. A. Zourmand, H.N. Ting, S.M. Mirhassani, Gender classification in children based on speech characteristics: using fundamental and formant frequencies of Malay vowels. J. Voice **27**(2), 201–209 (2013)