



Closed-set automatic speaker identification using multi-scale recurrent networks in non-native children

Kodali Radha¹ · Mohan Bansal¹

Received: 10 August 2022 / Accepted: 4 March 2023 / Published online: 18 March 2023

© The Author(s), under exclusive licence to Bharati Vidyapeeth's Institute of Computer Applications and Management 2023

Abstract Children may benefit from automatic speaker identification in a variety of applications, including child security, safety, and education. The key focus of this study is to develop a closed-set child speaker identification system for non-native speakers of English in both text-dependent and text-independent speech tasks in order to track how the speaker's fluency affects the system. The multi-scale wavelet scattering transform is used to compensate for concerns like the loss of high-frequency information caused by the most widely used mel frequency cepstral coefficients feature extractor. The proposed large-scale speaker identification system succeeds well by employing wavelet scattered Bi-LSTM. While this procedure is used to identify non-native children in multiple classes, average values of accuracy, precision, recall, and F-measure are being used to assess the performance of the model in text-independent and text-dependent tasks, which outperforms the existing models.

Keywords Automatic speaker identification · Non-native children speech · Text-dependent speech · Text-independent speech · Wavelet scattering transform · Bi-LSTM · RNN

1 Introduction

Automatic speaker recognition (ASR) is the approach used to automatically authenticate the identity of speaker by exploiting speaker-specific characteristics included within the speech waves, allowing accessibility for various applications [1, 2]. The speaker recognition is categorized into two sub-areas: speaker identification and speaker verification. The technique of identifying which of the enrolled speakers delivered a given utterance is known as speaker identification. The technique of admitting or refusing a speaker's reported identity is known as speaker verification. Furthermore, depending on the type of speech, speaker identification systems are classified as: text-dependent (defined passwords) and text-independent (undefined passwords). Text-dependent speaker identification (TD-SI) models are typically template-based or sequence-matching strategies in which the recognition passages are preset or available in advance [3]. However, preset phrases aren't really allowed in diverse applications, notably forensic investigations and surveillance. Furthermore, regardless of the subject of the speech, humans can distinguish speakers. Therefore, text-independent speaker identification (TI-SI) models have gained attention, as it is more difficult to imitate an unknown utterance [4]. In fact, social media platforms are the most popular among teens and young adults, with over half of youngsters aged 8–17 are using the internet and maintaining accounts on social media website [5]. However, several speech technology applications face difficulty in identifying non-native children due to various reasons of child's immaturity in the areas of cognition, phonology, and physiological development (vocal tract) [6, 7]. In particular, the pitch of children's voice is higher, and perceptual elements like formants appear at higher frequencies [8]. Despite, significant research focuses on automated identification of adult

✉ 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

speakers, concerns and procedures relevant to speaker identification in non-native children is still elusive and exposed to numerous exterior risks.

The broadly utilized feature engineering techniques in speech applications include conventional mel frequency cepstral coefficients (MFCC) and linear predictive cepstral coefficients (LPCC) [9]. A discrete cosine transform (DCT) is responsible in the back-end to minimize redundancy in the speech signal and dividing them into feature coefficients with small dimensions, however, it has a limitation of fixed resolution. Contrarily, the loss of high-frequency information caused by these feature extraction techniques is compensated by the multi-resolution (multi-scale) offered by the wavelet scattering transform (WST) [10]. The key contribution of this study is to identify the non-native child speaker in both text-dependent (TD) and text-independent (TI) speech tasks particularly in multiple classes, because as population grows larger, the performance of the model drops.

The details of the article are as follows: Sect. 2 provides a comprehensive insight into recent state-of-the-art models in children speaker identification. The proposed framework is presented in Sect. 3 and includes details of the speech corpus, wavelet scattered feature extraction, and Bi-LSTM model in multiple class classification. The details of the experimental setup, evaluation metrics, and reported outcomes of the non-native child speaker identification in both TD and TI speech tasks are given in Sect. 4. Lastly, Sect. 5 concludes with an expansion of future work.

2 Related work

Humans consider a variety of cues to identify the speaker, notably voice pitch, a particular laughter, and frequently used words [11]. However, the speaker identification system that has been built so far relies on short-term characteristics of the speech spectrum. Despite the fact that a lot of progress has indeed been made on para-linguistic analysis (age, gender, language, nativity, emotion, identity) of adult speaker identification [12–14]. It is astonishing that, there is comparatively less research done on children's speech in this field because para-linguistic analysis of children has a lot of significant challenges. They include high spectral and temporal randomness due to “the onset of puberty” [15], immature vocal tracts, the development of fine motor skills during language acquisition, and high fundamental frequencies than those of adults [16]. Children's speech technology is regulated by nativity and accent, as distinguished by para-linguistic characteristics [17]. From literature, modern speaker identification systems commonly deploy MFCC

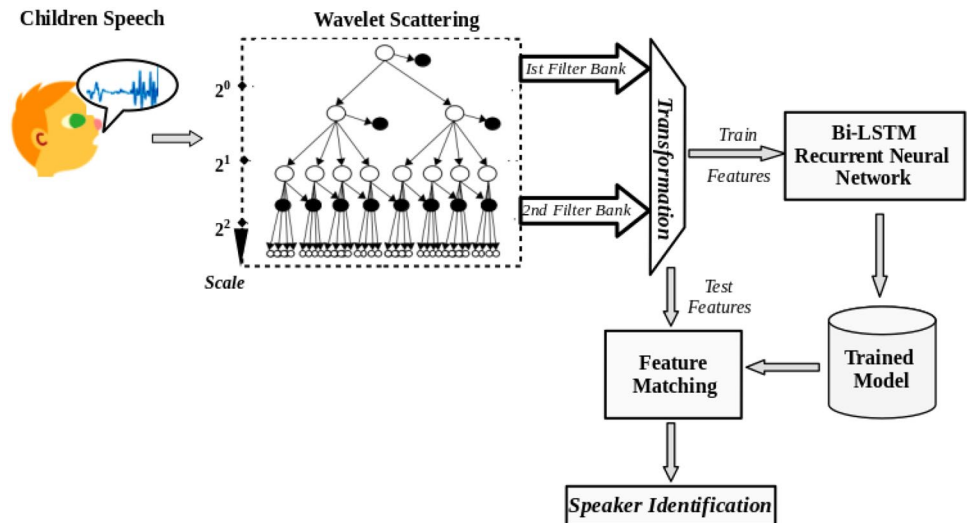
features trained with Gaussian mixture model (GMM) based methods to identify the speaker [2]. This well-known study was used in speaker recognition for children's speech, with OGI kids corpus trained on Gaussian mixture model-universal background model (GMM-UBM) and Gaussian mixture model-support vector machine (GMM-SVM) models [18] to obtain an accuracy of 90–99% for children between the age of 5 and 16. Additionally, employing GMM-UBM systems, the study on children's speech has continued to identify child speakers by gender and age group [19]. Recently, a work presented on children automatic speaker verification (ASV) in low resource condition using i-vector, x-vector and probabilistic linear discriminant analysis harnessed with audio augmentation techniques [20]. The success of wavelet scattered feature extraction integrated with a convolution neural network (CNN) for end-to-end adult speaker identification [21, 22] inspired us to apply the scattering network integrated with recurrent neural network (RNN) to the suggested non-native child speaker identification system. Wavelet scattering network (WSN) is used as a feature extractor in this study on TI-SI, overcoming the limitation of MFCCs. A typical approach found that, using a small amount of TI speech data to train Bi-LSTM was successful in large-scale speaker identification system [23].

Although speaker identification systems are effective in situations where there are few classes, as the number of classes grows, it becomes more difficult to identify the speaker, particularly in children who might not be native speakers. Therefore, the main objective of this work is to develop accurate methods for identifying non-native child speakers using deep RNNs and to evaluate the significance of information in filter coefficients of wavelet scattering networks.

3 Proposed system overview

This section presents a comprehensive procedure for identifying non-native child speakers across multiple classes in TD and TI speech tasks, making it the first article to use wavelet scattering transforms in children's speech. Raw audio files are first transformed into wavelet feature coefficients of order 2 using the wavelet scattering network, which serves as a feature extractor. A Bi-LSTM model of RNN is then trained on these features to solve sequential problems and acquire long-term dependence on inputs [24]. After training the Bi-LSTM model on the wavelet feature coefficients extracted from the raw audio files, the model can be used to predict the speakers in a separate test set, as shown in Algorithm 1. The process of predicting the speakers involves matching the test features with the trained features as shown

Fig. 1 Proposed child speaker identification system using multi-scale recurrent neural networks



in Fig. 1. The trained features are first fed into the Bi-LSTM model to obtain the corresponding sequence of hidden states, which are then used to predict the speakers using a softmax classifier. The test feature is then assigned to the class with the highest probability of the predicted speakers, based on the softmax output of the Bi-LSTM model.

Algorithm 1 Closed-Set Automatic Speaker Identification in Non-Native Children

Input: Text-dependent & Text-independent non-native children speech data(S) [Train 80%, Test 20%]

Feature Extraction from Raw Data:

Initialize: Scattering paths (n), time windows, invariance scale

for $n = 1:m$ **do**

$W_I[\lambda_1, \lambda_2, \dots, \lambda_n]S = ||S * \varphi_{\lambda_1} | * \varphi_{\lambda_2} | \dots \varphi_{\lambda_n} | * \rho_I(t)$
 where W_I denotes wavelet coefficients of m layers, S is the input speech, λ denotes the center frequencies of band pass filter, $\rho_I(t)$ is the averaging low-pass filter and $*$ indicates the convolution operator

end

for each epoch do

Generate hidden state ($h^{(t)}$), memory cell ($C^{(t)}$) via Bi-LSTM Model with the current input ($x^{(t)}$) & previous hidden state ($h^{(t-1)}$)

$\vec{h}^{(t)} = \tanh(U_{\vec{h}}[\vec{h}^{(t-1)}, x^{(t)}] + b_{\vec{h}})$

▷ Forward Pass of Bi-LSTM Model

$\overleftarrow{h}^{(t)} = \tanh(U_{\overleftarrow{h}}[\overleftarrow{h}^{(t-1)}, x^{(t)}] + b_{\overleftarrow{h}})$

▷ Backward Pass of Bi-LSTM Model

end

Generate classification output $y^{(t)}$ via fully connected layer with input $x^{(t)}$

Result: Child speaker identification in 20 classes [F_k, M_l], F denotes female class and M denotes male class, where $k = 1, 2, \dots, 11$ & $l = 1, 2, \dots, 9$.

3.1 Dataset

A total of 20 children, 11 females and 9 males aged 7–12 years are considered for data collection. All of the children are non-native English speakers who speak native Telugu, an Indian regional language, in addition to English. The data comprises two types of speech: TD speech and TI speech. The children repeat the preset phrases presented on

the screen in the TD speech activity (e.g: “Everyone in my school likes icecream”). In the TI speech task, children narrate a picture (e.g: pencil) on the screen in their own words (e.g: “This is a pencil. I will write my homework”).

The dataset consists around 3.32 h of children’s speech with 2898 recordings, 2173 from the TD speech task and 725 from the TI speech task. It is evident that TI speech data is less than TD speech data because children are more comfortable reading the text presented on screen rather than narrating a picture during data collection. This is because non-native children are speaking English as a medium of instruction, and the emergence of Covid-19 has severely hampered children’s learning and development in L2 proficiency [25]. All of the audio files are in .wav standard and use a stereo channel at 44.1KHz sampling rate and 16 bits per sample. Environmental settings such as door closing, bell ringing, and other children chatting in adjacent classrooms are also included because the audio recordings of non-native children were acquired during class days. Detailed description of database is provided at Non-Native Children Speech Mini Corpus [26].

3.2 Wavelet scattering transform

In natural settings, speech signals produced by children often demonstrate gradual changes or fluctuations that are occasionally interrupted by abrupt transients [27]. In such instances, wavelet scattering transform (WST) is employed as shown in Eq. (1), which may reflect the local features of signals in both the time and frequency domains providing multi-resolution [28, 29]. The WST is simply a bunch of convolutions with a series of band-pass filters (wavelet filter bank), non-linearity (modulus), and pooling [30]. Each layer must be capable of performing linear and non-linear

operations. Initially, in each layer, the convolution linear operation of the specified complex wavelet filters with the input signal is accomplished, followed by the modulus non-linearity is applied. Finally, a low-pass filter is used to determine the local average or pooling.

$$S_W(\alpha, \beta) = \frac{1}{\sqrt{\alpha}} \int_{-\infty}^{\infty} \hat{\varphi}\left(\frac{t-\beta}{\alpha}\right) s(t) dt \tag{1}$$

where $S_W(\alpha, \beta)$ is WST of $s(t)$ which is decomposed into its basis functions with α be the binary or dyadic scale and β be the binary or dyadic shift and $\hat{\varphi}(\cdot)$ denotes complex conjugate of wavelet function [31].

The first-order scattering coefficients are mathematically written as Eq. (2):

$$W_I[\lambda_1]S = |S * \varphi_{\lambda_1}| * \rho_I(t) \tag{2}$$

where S is the input speech, $\varphi_{\lambda_1}(t)$ denotes dilated wavelet at λ_1 center frequency of first order bandpass filter, $\rho_I(t)$ is the averaging low-pass filter and $*$ indicates the convolution operator. The high-frequency coefficients are obtained by local averaging by $\rho_I(t)$ to maintain better stability, however, this results in a loss of high-frequency data. The wavelet modulus is fragmented further in second-order scattering as shown in Eq. (3) and in order to preserve high-frequency information.

$$W_I[\lambda_1, \lambda_2]S = ||S * \varphi_{\lambda_1}| * \varphi_{\lambda_2}| * \rho_I(t) \tag{3}$$

According to the first-order coefficients, these coefficients are averaged using a low-pass filter ρ_I of length T , which provides local invariance to time shifts.

Similarly, in the m^{th} layer, the sequence of n pre-defined wavelet paths $\varphi_{\lambda_1}, \varphi_{\lambda_2} \dots \varphi_{\lambda_m}$ are convoluted using coefficients from the $(m - 1)^{th}$ layer, which correspond to a prior wavelet $\varphi_{\lambda_{m-1}}$. Figure 2 depicts an example of WST for input speech with the scattering order, $M=4$ and scale, i is 4, and $m = 0, m = 1, m = 2,$ and $m = 3$ represents the filter bank order.

3.3 Bidirectional long-short term memory (Bi-LSTM)

Recurrent neural networks (RNN) were shown to be extremely effective at solving sequential problems [32, 33]. Long short-term memory (LSTM) networks were found in 1997 by Hochreiter and Schmidhuber [34] as an extension of RNN. There is rich literature available on LSTM due to the ability to learn long-term dependence of inputs by simply memorizing information over a long period of time and also competent in dealing with the vanishing/exploding gradient problem [35]. This is possible by incorporating an input gate ($\hat{g}_i^{(t)}$), an output gate ($\hat{g}_o^{(t)}$), a forget gate ($\hat{g}_f^{(t)}$), and a memory cell ($C^{(t)}$) in the LSTM network [36], as shown in the Fig. 3a. The memory cell is made up of two layers: the *sigmoid* layer, which is responsible for updating weights, and the *tanh* layer, which is responsible for formulating new cell state vectors ($\hat{C}^{(t)}$), that add to memory cells.

Bidirectional LSTM (Bi-LSTM), is an updated version of LSTM model that employ two LSTMs on the input data [37]. The proposed network learns not only from input cell ($C^{(t-1)}$) to the output cell ($C^{(t)}$) in the forward direction using Eq. (4), but also from output cell to input cell in the backward direction using Eq. (5), via two different hidden layers [38], in order to preserve previous and future information as depicted in Fig. 3b.

$$\vec{h}^{(t)} = \tanh(U_h^+[\vec{h}^{(t-1)}, x^{(t)}] + b_h^+) \tag{4}$$

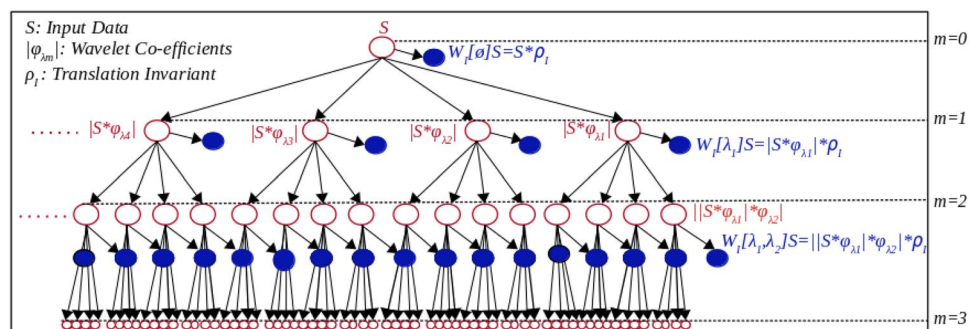
$$\overleftarrow{h}^{(t)} = \tanh(U_h^-[\overleftarrow{h}^{(t-1)}, x^{(t)}] + b_h^-) \tag{5}$$

Here $x^{(t)}$ denotes current input, $\vec{h}^{(t)}$ is the forward hidden sequences from $t = 1$ to T , and $\overleftarrow{h}^{(t)}$ is the backward hidden sequence from $t = T$ to 1, U_h^+ & U_h^- are the corresponding weights with bias vectors b_h^+ and b_h^- .

Finally, the output $y^{(t)}$, is obtained by appending the two hidden states with an output bias vector b_y as shown in Eq. (6).

$$y^{(t)} = U_h^+ \vec{h}^{(t)} + U_h^- \overleftarrow{h}^{(t)} + b_y \tag{6}$$

Fig. 2 Decomposition using wavelet scattering network. Red color denotes non-linear modulus after convolution with series of band-pass filter at center frequency (λ) & blue colour denotes outputs obtained via low-pass filter (ρ_I)



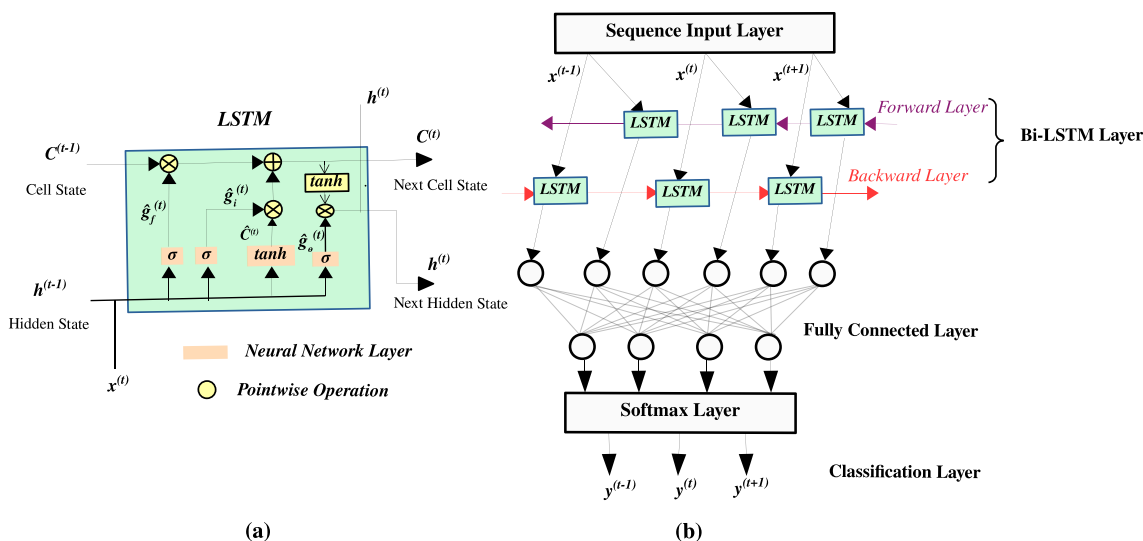


Fig. 3 a Basic details of LSTM cell. b Recurrent neural network with Bi-LSTM layer

Table 1 Details of train and test split in non-native children speech corpus

Type of children’s speech	Split	No. of utterances	Total
TD Speech	Train	1736	2173
	Test	437	
TI Speech	Train	580	725
	Test	145	

Implementing the gated cell (LSTM) twice enhances the ability to learn long-term dependencies and, as a result, improves prediction performance. A more detailed description of Bi-LSTM network configuration is discussed in Sect. 4.3.

4 Results and discussions

This section describes the experimental settings and outcomes in developing the child speaker identification model. Initially, the datasets were randomly split into train and test sets both in TD and TI speech. To gauge the overall system performance, average values of accuracy, precision, and recall metrics were employed and compared with the existing work. MATLAB R2022a was used for all investigations of children’s speech, including feature extraction, modeling, and identification.

4.1 Data processing

This article incorporates, a closed-set child speaker identification system that integrates TD and TI speech to promote voice biometric applications [39–41]. The overall corpus length is 3.34 h (~200 min) and comprises 2898 utterances, of which 2173 are TD speech data and 725 are TI speech data. Furthermore, as indicated in Table 1, each type of speech is randomly divided into 80% and 20% as training and testing data, respectively.

4.2 Feature extraction

Feature extraction is an important process in any recognition system since it provides the system to reliably identify speakers by transforming the raw data into feature vectors. The most popular feature extraction methods used in automatic speaker recognition systems are the MFCC and LPCC [42]. Filters with nonlinearly variable bandwidths are imposed over the STFT in MFCCs. Filter implementation may be considered as frequency domain pooling that negates the effect of instability. Unfortunately, this frequency pooling entails loss of signal information at higher frequencies [43]. To counteract the loss of high frequency information in MFCC, the WST approximates second order coefficients of the speech signal.

In the proposed speaker identification models, WST approach employs the Gabor (analytic Morlet) wavelet [31] was generated to a depth of 2. The very first layer comprised with eight Gabor wavelets per octave, whereas

Table 2 Training parameters of Adam optimizer

Parameter	Value
Gradient decay rate	0.9
Squared gradient decay rate	0.999
Epsilon (ϵ)	$1e^{-8}$
Initial learning rate	$1e^{-4}$
Drop rate	0.1
Drop period	10
L2 regularization	$1e^{-4}$
Gardient threshold	L2-norm
Maximum epochs	50
Mini-batch size	128
Shuffle	Every epoch

the second included one Morlet wavelet per octave and represented as quality factors $Q = [8 \ 1]$. The wavelets use a low pass scaling function to produce interpretations of children’s speech with low variation. The time scale for invariance is set at 500 ms and a batch size of 64. The coefficients are therefore standardized and log-transformed. As a result, analyzing children speech by employing the first and second orders filter banks of the WST extends the Mel filter bank representation, while conserving the information. Finally, the scattering network obtained has 411 paths by 25 scattering time windows for each audio signal.

4.3 Training network configuration

The proposed child speaker identification system was developed and trained using the Bi-LSTM model classifier. The training network comprises of five layers, with the extracted scattering coefficients from the wavelet scattering network served into a 5-layer RNN. The input layer is a sequence

layer that has 410 dimensions, followed by a Bi-LSTM layer with 512 hidden units and a 20-fully connected layer. The softmax function is required as the activation function in the output layer, which predicts a multinomial probability [44] for multiple speaker. Finally, a cross entropy is employed in the classification layer to update the network via computing the difference between the output and target multinomial probabilities is as shown in Fig. 3b. The proposed model was trained using “Adam (adaptive moment estimation)” [45], an optimization algorithm with an adaptive learning rate that evaluates individual learning rate for various parameters. The details of learning options used in both TD and TI speaker identification system are depicted in Table 2.

4.4 Evaluation metrics

The performance measures considered to evaluate the proposed non-native children identification system in TD and TI tasks are average accuracy, average precision, average recall, and average F1 score [46] as shown in Table 3, since this approach is employed to recognize children in multi-class problem [47]. Moreover, area under the receiver operator characteristic curve (AUROC) or ROC index is mostly used to measure the model’s scoring ability. This is a likelihood graph that effectively separates the signal from the noise by analyzing the true positive rate (TPR) to the false positive rate (FPR) at various threshold levels [48].

4.5 Text-dependent speaker identification (TD-SI)

To investigate the performance of speaker identification system in TD speech task of non-native children, Bi-LSTM network is trained on 1736 utterances and tested with 437

Table 3 Evaluation metrics in multiple class (J) speaker identification system

Metric	Formula	Evaluation procedure
Macro avg. accuracy	$\bar{A}_v = \frac{\sum_{j=1}^J \frac{TP_j + TN_j}{TP_j + FN_j + FP_j + TN_j}}{J}$	Basic arithmetic mean of correct predictions across all observations
Macro avg. precision	$\bar{P} = \frac{\sum_{j=1}^J \frac{TP_j}{TP_j + FP_j}}{J}$	Basic arithmetic mean of which determines the number of correct positive predictions from all positive predictions in imbalanced datasets
Macro avg. recall	$\bar{R} = \frac{\sum_{j=1}^J \frac{TP_j}{TP_j + FN_j}}{J}$	Compute the basic arithmetic mean of missed positive predictions
Macro avg. F-measure	$\bar{F1}\text{-Score} = 2 \times \frac{\bar{P}\bar{R}}{\bar{P} + \bar{R}}$ $= \frac{\sum_{j=1}^J \frac{TP_j}{TP_j + \frac{1}{2}(FP_j + FN_j)}}{J}$	Compute the basic harmonic mean of precision and recall

utterances from 11 female children and 9 male children. The training method consists of 13 iterations per epoch for 50 epochs, for a total of 650 iterations all through the training cycle. The TD-SI system achieves an overall average accuracy of 97.94% with a misclassification of 2.1%, primarily owing to the F6 speaker, who has a similar voice to F5 and F2 due to age resemblance. Male children M6 and M9 are often misclassified as female child F6 because of the fact that, occasionally male children whose vocals have not changed as a result of puberty may attain a close match with the females [15]. Overall average accuracy may not provide a deep insight into system performance because of the imbalance in the dataset in recognizing non-native speakers; consequently, the F-measure is used as a combined measure (Harmonic Mean) of precision and recall. In TD speech task, the F-measure of non-native speaker identification is 98.01%, with a respective precision of 98.02% and a recall of 98.01%, as displayed in the confusion matrix in Fig. 4.

Table 4 Performance measure of speaker identification system in terms of average accuracy (\bar{A}_v), average precision (\bar{P}), average recall (\bar{R}), average F-measure ($\bar{F}1$ -score)

Model	\bar{A}_v (%)	\bar{P} (%)	\bar{R} (%)	$\bar{F}1$ -Score(%)
TD-SI	97.94	98.02	98.01	98.01
TI-SI	96.6	97.32	96.94	97.1

Figure 6a and Fig. 6b show the ROC curves for the female and male classes in TD-SI tasks, respectively.

4.6 Text-independent speaker identification (TI-SI)

The reliability of the speaker identification model in TI speech of non-native children is shown in Table 4. The Bi-LSTM network is trained on 580 utterances of children aged 7-12 years and evaluated against 145 imbalanced

Fig. 4 Confusion matrix of text-dependent speaker identification system



Fig. 5 Confusion matrix of text-independent speaker identification system

		Confusion Matrix																				
		F1	F10	F11	F2	F3	F4	F5	F6	F7	F8	F9	M1	M2	M3	M4	M5	M6	M7	M8	M9	
Output Class	F1	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100%
		5.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	F10	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100%
		0.0%	4.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	F11	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100%
		0.0%	0.0%	5.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	F2	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100%
		0.0%	0.0%	0.0%	2.8%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	F3	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100%
		0.0%	0.0%	0.0%	0.0%	2.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	F4	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100%
		0.0%	0.0%	0.0%	0.0%	0.0%	5.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	F5	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	100%
		0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	6.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	F6	0	0	0	0	0	0	0	7	0	0	1	0	0	0	0	0	0	0	0	0	87.5%
		0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	4.8%	0.0%	0.0%	0.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	12.5%
	F7	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	100%
		0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	4.8%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	F8	0	2	0	0	0	0	0	0	0	0	7	1	0	0	0	0	0	0	0	0	70.0%
	0.0%	1.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	4.8%	0.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	30.0%	
F9	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	100%	
	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	4.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	
M1	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	100%	
	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	2.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	
M2	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	100%	
	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	5.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	
M3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	100%	
	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	5.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	
M4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	100%	
	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	5.5%	0.0%	0.0%	0.0%	0.0%	0.0%	
M5	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	8	0	0	0	88.9%	
	0.0%	0.0%	0.0%	0.0%	0.0%	0.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	5.5%	0.0%	0.0%	0.0%	11.1%	
M6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	100%	
	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	5.5%	0.0%	0.0%	0.0%	
M7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	100%	
	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	6.2%	0.0%	0.0%	
M8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	100%	
	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	5.5%	0.0%	
M9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	100%
	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	4.8%	0.0%
	100%	75.0%	100%	100%	100%	88.9%	100%	100%	100%	100%	75.0%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	96.6%
	0.0%	25.0%	0.0%	0.0%	0.0%	11.1%	0.0%	0.0%	0.0%	0.0%	25.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	3.4%
	F1	F10	F11	F2	F3	F4	F5	F6	F7	F8	F9	M1	M2	M3	M4	M5	M6	M7	M8	M9		

utterances from 20 speakers. The training procedure involves 4 iterations per epoch for 50 epochs, results in maximum of 200 iterations in the training cycle. The TI-SI system achieves an overall average accuracy of 96.6% with an error rate of 3.4%. Two F10’s utterances are incorrectly identified as F8, one F4’s utterance is incorrectly anticipated as M5’s, and two F9’s utterances are poorly predicted as F6’s and F8’s, respectively. The above misclassification is regarded to be largely from female children aged 7 and 8, because the females in this age group have comparable pitch. As this TI-SI model includes limited test data, perhaps a misclassification of one utterance results in a considerable loss, as illustrated in the confusion matrix in Fig. 5. The ROC curves for the female and male classes in TI-SI tasks, is depicted in Fig. 6c and 6d respectively. Finally, the F-measure is also employed as a consolidated unit of precision and recall. In TI speech task, the F-measure of non-native speaker identification is 97.1%, with a respective precision of 97.32% and a recall of 96.94%.

4.7 Comparative analysis of proposed system with earlier work

This section provides a much more detailed comparison of current state-of-the-art models for identifying child speakers, including information on the speech corpus type, feature extraction method, modeling strategy, and accuracy results. Additionally, adult speaker identification systems using wavelet scattered features are also compared. The proposed model outperforms earlier approaches in children’s speech as depicted in Table 5 in both TD and TI speech tasks.

5 Conclusion and future work

In this study, non-native child speaker identification systems were built using WSTs as feature vectors, which directly acquired speaker discriminatory information from raw speech signals. A corpus of non-native children learning English as a second language is being collected and examined in natural environmental settings in order

Fig. 6 a, b Represents ROC curves for text-dependent speaker identification. c, d Represents ROC curves for text-independent speaker identification

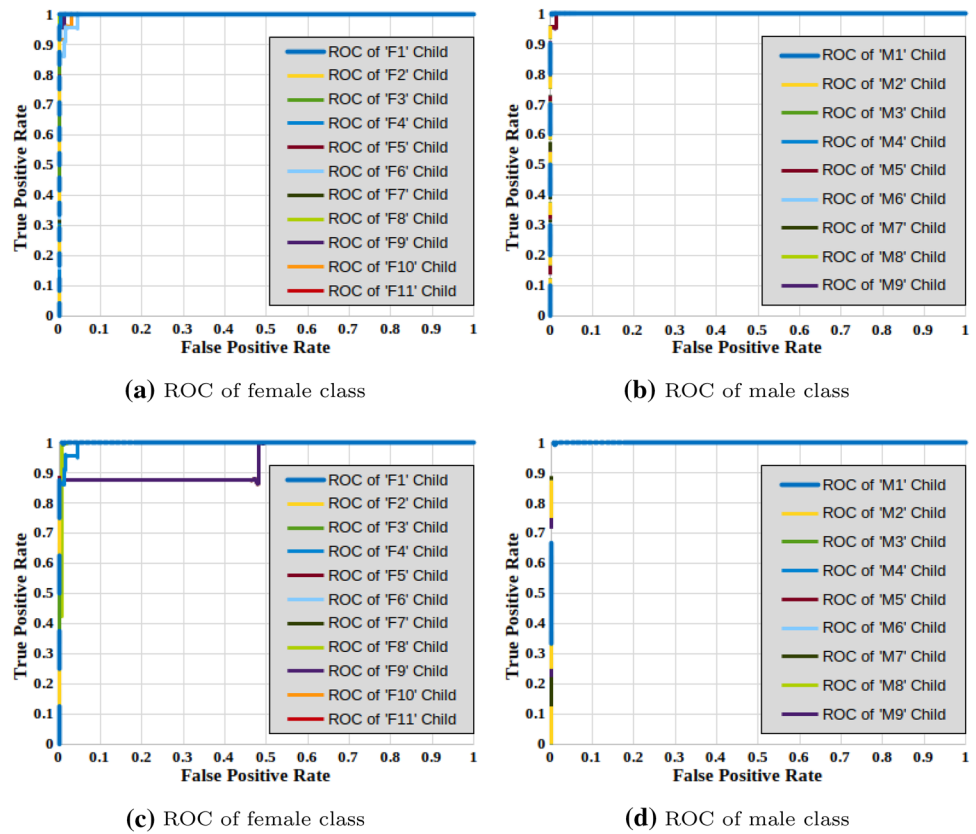


Table 5 Comparative analysis of earlier state-of-the-art results in speaker identification

Year/author	Corpus type	Fronte-end feature extraction	Modeling approach	Performance
2012/Safavi et al. [49]	OGI Kids corpus of text dependent and text independent speech	Isolated and full sub-bands MFCC+ Δ + $\Delta\Delta$	GMM-UBM and GMM-SVM	90%, 96% and 99% for 3 age groups respectively (Speaker recognition system alone)
2018/Safavi et al. [19]	OGI Kids corpus of text dependent and text independent speech	Isolated and full sub-bands MFCC+ Δ + $\Delta\Delta$	GMM-UBM and GMM-SVM and i-vectors	78.16%, 91.13% and 95.38% for 3 age groups respectively (automatic speaker, age, gender, identification system)
2019/Jia et al. [50]	Text dependent children speech	Multi-dimensional features like prosody, sound quality and spectral based features	TDNN and Bi-LSTM	Automatic speaker, age, gender, identification system, but high accuracy for speaker recognition
2020/Ghezaiel et al. [21]	Adult speech in TIMIT and Librispeech datasets text independent speech	MFCC, wavelet scattering, FBANK	DNN, CNN, SincNet	79% and 88% in short utterances of 2 s and 4 s respectively
2021/Ghezaiel et al. [22]	Adult speech in Librispeech and TIMIT datasets text independent speech	Wavelet scattering	CNN	99.2% & 98.1% for both dataset respectively
Proposed work	Children text dependent and text independent speech	Wavelet scattering transform	Bi-LSTM	Overall accuracy of 97.9% and 96.6% for text dependent & text independent speech respectively

to benefit children’s speech technology. The proposed system proved to work reasonably well for both TD and TI speech scenarios in identifying over 20 unique children. The experimental findings showed that both the TD-SI

and TI-SI models performed almost equally well with an overall accuracy of 97.9% and 96.6% respectively using scatter features trained on Bi-LSTM RNNs. The outcome of the study shows that the suggested closed set speaker

identification method is reliable and efficient. The significant findings indicate that this system may be deployed to a variety of tasks, including control and security access. Future developments of the proposed study will include the automatic identification of speaker, nativity, age, and gender for a large corpus of non-native children's speech.

Acknowledgements The authors extend their thanks to the High-Performance Computing Lab at VIT-AP University for supplying the resources and assistance that made this research possible. The access to state-of-the-art computing equipment and expert technical support greatly aided in completing this article.

Data Availability The open access data that support the findings of this study is made publicly available in KAGGLE repository, <https://doi.org/10.34740/KAGGLE/DS/2160743>. The more details about the data collection is given in Sect. 3.1.

References

- Furui S (1996) An overview of speaker recognition technology. *Autom Speech Speaker Recognit* 31–56
- Bai Z, Zhang X-L (2021) Speaker recognition based on deep learning: An overview. *Neural Netw* 140:65–99
- Kinnunen T, Li H (2010) An overview of text-independent speaker recognition: From features to supervectors. *Speech Commun* 52(1):12–40
- Reynolds DA (2002) An overview of automatic speaker recognition technology. In: 2002 IEEE international conference on acoustics, speech, and signal processing, vol 4. IEEE, p 4072
- Badillo-Urquiola K, Smriti D, McNally B, Golub E, Bonsignore E, Wisniewski PJ (2019) 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
- Gretter R, Matassoni M, Falavigna GD, Keelan E, Leong CW (2020) Overview of the interspeech TLT2020 shared task onsr for non-native children's speech. In: *Interspeech 2020*, pp 245–249
- Radha K, Bansal M (2022) Audio augmentation for non-native children's speech recognition through discriminative learning. *Entropy* 24(10):1490
- Lee S, Potamianos A, Narayanan S (1999) Acoustics of children's speech: developmental changes of temporal and spectral parameters. *J Acoust Soc Am* 105(3):1455–1468
- Kumar A, Mittal V (2021) Hindi speech recognition in noisy environment using hybrid technique. *Int J Inf Technol* 13(2):483–492
- Kalra D, Gupta R (2021) Novel model for pitch estimation using hybrid DWT-DCT HPS. *Int J Inf Technol* 13(4):1409–1415
- Schmidt-Nielsen A, Crystal TH (2000) Speaker verification by human listeners: experiments comparing human and machine performance using the NIST 1998 speaker evaluation data. *Dig Signal Process* 10(1–3):249–266
- Schuller B, Steidl S, Batliner A, Burkhardt F, Devillers L, Müller C, Narayanan S (2013) Paralinguistics in speech and language-state-of-the-art and the challenge. *Comput Speech Lang* 27(1):4–39
- Bansal M, Sircar P (2019) Phoneme based model for gender identification and adult-child classification. In: 2019 13th international conference on signal processing and communication systems (ICSPCS). IEEE, pp 1–7
- Schuller B, Batliner A (2013) Computational paralinguistics: emotion, affect and personality in speech and language processing, pp 1–321
- Rogol AD, Clark PA, Roemmich JN (2000) Growth and pubertal development in children and adolescents: effects of diet and physical activity. *Am J Clin Nutr* 72(2):521–528
- Gerosa M, Giuliani D, Brugnara F (2007) Acoustic variability and automatic recognition of children's speech. *Speech Commun* 49(10–11):847–860
- Radha K, Bansal M, Shabber SM (2022) Accent classification of native and non-native children using harmonic pitch. In: 2022 2nd international conference on artificial intelligence and signal processing (AISP). IEEE, pp 1–6
- Safavi S, Najafian M, Hanani A, Russell MJ, Jancovic P, Carey MJ (2016) Speaker Recognition for Children's speech. <https://doi.org/10.48550/ARXIV.1609.07498>
- Safavi S, Russell M, Jančovič P (2018) Automatic speaker, age-group and gender identification from children's speech. *Comput Speech Lang* 50:141–156
- Shahnawazuddin S, Ahmad W, Adiga N, Kumar A (2021) Children's speaker verification in low and zero resource conditions. *Dig Signal Process* 116:103115
- Ghezaiel W, Luc B, Lézoray O (2020) Wavelet scattering transform and CNN for closed set speaker identification. In: 2020 IEEE 22nd international workshop on multimedia signal processing (MMSp). IEEE, pp 1–6
- Ghezaiel W, Brun L, Lézoray O (2021) Hybrid network for end-to-end text-independent speaker identification. In: 2020 25th international conference on pattern recognition (ICPR). IEEE, pp 2352–2359
- Nammous MK, Saeed K, Kobojeck P (2022) Using a small amount of text-independent speech data for a Bi-LSTM large-scale speaker identification approach. *J King Saud Univ Comput Inf Sci* 34(3):764–770
- Su Y, Kuo C-CJ (2019) On extended long short-term memory and dependent bidirectional recurrent neural network. *Neurocomputing* 356:151–161
- Basri S, Hawaldar IT, Nayak R, Rahiman HU (2022) Do academic stress, burnout and problematic internet use affect perceived learning? evidence from India during the covid-19 pandemic. *Sustainability* 14(3):1409
- Radha K, Bansal M (2022) Non-native children speech mini corpus. Kaggle. <https://doi.org/10.34740/KAGGLE/DS/2160743>
- Bijou SW, Peterson RF, Harris FR, Allen KE, Johnston MS (1969) Methodology for experimental studies of young children in natural settings. *Psychol Rec* 19:177–210
- Mateo C, Talavera JA (2020) Bridging the gap between the short-time Fourier transform (STFT), wavelets, the constant-Q transform and multi-resolution STFT. *Signal Image Video Process* 14(8):1535–1543
- Sifuzzaman M, Islam MR, Ali M (2009) Application of wavelet transform and its advantages compared to Fourier transform. *J Phys Sci* 13:121–134
- Bruna J, Mallat S (2013) Invariant scattering convolution networks. *IEEE Trans Pattern Anal Mach Intell* 35(8):1872–1886
- Vetterli M, Herley C (1992) Wavelets and filter banks: theory and design. *IEEE Trans Signal Process* 40(ARTICLE):2207–2232
- LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521(7553):436–444
- Jayanthi K, Mohan S et al (2022) An integrated framework for emotion recognition using speech and static images with deep classifier fusion approach. *Int J Inf Technol* 1–11
- Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Comput* 9(8):1735–1780

35. Staudemeyer RC, Morris ER (2019) Understanding LSTM—a tutorial into long short-term memory recurrent neural networks. [arXiv:1909.09586](https://arxiv.org/abs/1909.09586)
36. Yu Y, Si X, Hu C, Zhang J (2019) A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput* 31(7):1235–1270
37. Schuster M, Paliwal KK (1997) Bidirectional recurrent neural networks. *IEEE Trans Signal Process* 45(11):2673–2681
38. Singh B, Jaiswal R (2022) Automation of prediction system for temporal data. *Int J Inf Technol* 14(6):3165–3174
39. Punyani P, Gupta R, Kumar A (2022) A multimodal biometric system using match score and decision level fusion. *Int J Inf Technol* 14(2):725–730
40. Shashidhar R, Patilkulkarni S, Puneeth S (2022) Combining audio and visual speech recognition using LSTM and deep convolutional neural network. *Int J Inf Technol* 1–12
41. Ann OC, Theng LB (2011) Biometrics based assistive communication tool for children with special needs. In: 2011 7th international conference on information technology in Asia. IEEE, pp 1–6
42. Hansen JH, Hasan T (2015) Speaker recognition by machines and humans: a tutorial review. *IEEE Signal Process Mag* 32(6):74–99
43. Andén J, Mallat S (2014) Deep scattering spectrum. *IEEE Trans Signal Process* 62(16):4114–4128
44. Mohammed AA, Umaashankar V (2018) Effectiveness of hierarchical softmax in large scale classification tasks. In: 2018 international conference on advances in computing, communications and informatics (ICACCI). IEEE, pp 1090–1094
45. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. <https://doi.org/10.48550/ARXIV.1412.6980>
46. Grandini M, Bagli E, Visani G (2020) Metrics for multi-class classification: an overview. <https://doi.org/10.48550/ARXIV.2008.05756>
47. Kaur P, Gosain A (2018) Issues and challenges of class imbalance problem in classification. *Int J Inf Technol* 1–7
48. Güvenir HA, Kurtcephe M (2012) Ranking instances by maximizing the area under ROC curve. *IEEE Trans Knowl Data Eng* 25(10):2356–2366
49. Safavi S, Najafian M, Hanani A, Russell MJ, Jancovic P, Carey MJ (2016) Speaker recognition for children’s speech. [arXiv:1609.07498](https://arxiv.org/abs/1609.07498)
50. Jia N, Zheng C, Sun W (2019) Children’s speaker recognition method based on multi-dimensional features. In: International conference on advanced data mining and applications. Springer, pp 462–473

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.