International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Comparative Analysis of Discrete Wavelet Transforms and Deep Sequence Classifiers for Robust Text-Independent Speaker Identification

Ajaykumar Vasantlal Patil

Department of Electronics and Communication Engineering, Government Polytechnic Vadnagar, Gujarat, India
Email: ajaypatil82@gmail.com

Abstract:
Speaker independent automated Speaker Identification (ASI) systems are dependent on techniques of acoustic feature extraction that preserve the speaker perception of some vocal tract characteristics of the speaker while minimising the effects of background noise, inter-session variation and phonetic differences [1] [2]. Mel-Frequency Cepstral Coefficients (MFCCs), which are the most commonly used traditional acoustic features obtained from the Short-Time Fourier Transform (STFT), have been found to be uniformly time-frequency windowed and the information lost in the higher frequencies [4, 11]. To overcome these dynamic limits, this research aims to make an empirical study about Multiresolution Discrete Wavelet Transforms (DWT) combined with statistical machine learning and deep sequential classification models for text independent speaker identification [5] and [9]. The three different families of mother wavelets are Daubechies (db4), Symlets (sym5), and Coiflets (coif3) with a decomposition structure of 5-level dyadic decomposition (D_{1}–D_{5} and A_{5}) [10, 12]. A statistical vector of 24 dimensions (logarithmic energy, variance, Shannon entropy and kurtosis) is derived from each frame's decomposed sub-bands [12], [13]. The classification abilities of these wavelet signatures are systematically compared by using Support Vector Machines with a Radial Basis Function kernel (SVM-RBF) [15], 1-D Convolutional Neural Networks (1D-CNN) [5, 14] and Bidirectional Long Short-Term Memory networks (Bi-LSTM) [4, 16]. The Daubechies (db4) wavelet paired with an SVM-RBF classifier performed very well in terms of identification accuracy (83.33%, compared to 73.46% for deep sequence models under the same conditions [6, 15]).

Keywords: Automatic Speaker Identification, Discrete Wavelet Transform, Multiresolution Analysis, Support Vector Machine, 1-D Convolutional Neural Network, Bidirectional LSTM, LibriSpeech.

1. Introduction

Automatic Speaker Recognition (ASR) technologies can be generally divided into two classes: Speaker Verification (SV), which involves an either/or decision in an identity test, and Speaker Identification (SI), which involves the identification of an unknown speaker from a closed or open candidate list [1]–[3]. Voice biometrics enables a convenient, easy-to-access, secure way to verify identity in telephony banking applications, forensic analysis, multi-user voice-assisted systems, and to establish personalized human computer interaction [2, 7, 8]. The automatic speaker identification (ASI) for the text

independent recognition (TI-ASI) is performed on arbitrary and unconstrained spoken input without the aid of particular trigger phrases [3, 11]. The acoustic extraction front-end, therefore, needs to exploit intrinsic anatomical signatures to reflect glottal excitation cycles, vocal tract geometry and formant resonances while simultaneously being invariant to spoken text and environmental noise [1], [11].

Conventional spectral representation is limited by two problems.There are two fundamental problems with the conventional spectral representation.

ASR technologies can be generally divided into two types: speaker verification, which is a binary decision making process to verify an identity claim, and speaker identification, where the speaker identity of an unknown speaker is identified from a closed or open set of candidates [1]–[3]. Voice biometrics offers a non-obtrusive and natural way of secure identity authentication in telephony banking, forensic analysis, multi-user voice assistants and personalized human-computer interaction [2], [7], [8]. In text independent automatic speaker identification (TI-ASI), a speaker must be identified by arbitrary, unconstrained speech input without using a particular ID phrase [3, 11]. Intrinsic anatomical signatures, such as glottal excitation cycles, vocal tract geometry and formant resonances, must thus be captured by the acoustic extraction front-end in a way that is invariant to the spoken text and the acoustic environment [1, 11].

1.2 Discrete Wavelet Transforms (DWT) and Multiresolution Analysis

Discrete Wavelet Transform (DWT) [9] [10] is one of the approaches that removes the limitations of short-time Fourier windowing. The DWT creates a dyadic octave-band filter tree by scaling and shifting a base mother wavelet ($\psi(t)$) [5], [6], [12]. This is a natural adaptation of the resolution: it can offer a high time resolution at high frequencies to capture the short transient of the sound; a high frequency resolution at low frequencies to resolve the periodicity and pitch harmonics of the voice.

The choice of the mother wavelet directly determines the approximation power, phase symmetry, and filter support of filters [5, 12]:

Daubechies (db4): Compactly supported orthogonal wavelets with maximal number of vanishing moments for a given support length, which allow for maximum isolation of abrupt speech transients in time [5].

Symlets (sym5): Wavelets designed to be least asymmetric and modified to impart near-linear phase responses in order to retain phase alignment throughout the harmonic formant frequencies [12].

Coiflets (coif3): Orthogonal wavelets with vanishing moments for both the wavelet function $\psi(t)$ and scaling function $\phi(t)$ which allow the accurate representation of polynomials over sustained vowel segments [13].

1.3 Machine Learning and Deep Sequence Classifiers

The downstream classifier has to build complex speaker decision boundaries after extracting multiresolution sub-band representations. This study investigates three basic classification methods:

Support Vector Machine with Radial Basis Function (SVM-RBF): A kernel-based statistical classifier based on Vapnik's statistical learning theory [15] maps input features into a high dimensional reproducing kernel Hilbert space to build optimum separating hyperplanes and is known to have a high generalization power for small-to-medium size dataset [3], [15].

1D Convolutional Neural Network (1D-CNN): A deep feed forward network that is able to learn local spatial correlation and acoustic patterns invariant to time [5], [6], [14].

Bidirectional Long Short-Term Memory (Bi-LSTM): Recurrent neural network with forward and backward temporal gate to model co-articulation dynamics and long-range sequential context over acoustic frames [4].

1.4 Principal Research Contributions

We create a unified 5-level DWT feature processing framework, which computes 4 statistical features (logarithmic energy, variance, Shannon entropy, and kurtosis) in 6 decomposed sub-bands to get a compact 24-dimensional feature vector for each analysis frame [12], [13].

Strictly standardized training and testing conditions are used and a unified benchmark evaluation is conducted for all 9 different combinations of mother wavelets (db4, sym5, coif3) and classification models (SVM-RBF, 1D-CNN, Bi-LSTM).

To avoid overfitting in multi-class configurations, we augment the data with a 3× scheme (add white noise with AWGN and scale the amplitude) on the 30-speaker LibriSpeech dev-clean dataset [17] to perform experiments.

We profile the computational trade-offs for each of the following: feature extraction runtime, model training latency for the CPU/GPU hardware, and real-time frame inference latency [6,14].

2. Mathematical Formulation and Theoretical Framework

2.1 Discrete Wavelet Transform and Sub-Band Partitioning

Multiresolution analysis decomposes a continuous discrete-time speech signal $s[n] \in L^2(R)$ through successive approximation spaces at decreasing scale resolutions [9], [10]. The decomposition relies on two scaling equations defining the scaling function $\phi(t)$ and mother wavelet $\psi(t)$:

$$\phi(t) = \sqrt{2} \sum_{n=-\infty}^{\infty} h\,[n]\phi(2t-n)$$

$$\psi(t) = \sqrt{2} \sum_{n=-\infty}^{\infty} g\,[n]\phi(2t-n)$$

where $h[n]$ denotes a low-pass reconstruction filter and $g[n] = (-1)^n h[1-n]$ forms the corresponding high-pass quadrature mirror filter [10]. At each decomposition stage $j \in \{1, 2, \dots, J\}$, the approximation coefficients $A_j[k]$ and detail coefficients $D_j[k]$ are obtained by filtering the previous approximation signal $A_{j-1}[m]$ followed by dyadic downsampling [5], [10]:

$$A_j[k] = \sum_{m=-\infty}^{\infty} A_{j-1}\,[m]\,h[2k-m]$$

$$D_j[k] = \sum_{m=-\infty}^{\infty} A_{j-1}\,[m]\,g[2k-m]$$

where $A_0[m] = s[m]$ represents the original speech signal. For audio sampled at $F_s = 16kHz$ decomposed across $J = 5$ levels, the Nyquist spectrum (0–8000 Hz) is split into five detail sub-bands (D_1 to D_5) and one terminal approximation sub-band (A_5), as outlined in Table 1 [4], [12].

Table 1: Frequency Sub-band Partitioning for 5-Level DWT (Fs = 16 kHz)

Level	Sub-band	Bandwidth (Hz)	Dominant Acoustic Correlates
Level 1	Detail D1	4000–8000	Unvoiced fricatives, plosive bursts, high-frequency transients
Level 2	Detail D2	2000–4000	Upper formants (F3, F4), consonant spectral cues
Level 3	Detail D3	1000–2000	Second vowel formant (F2), nasal resonances
Level 4	Detail D4	500–1000	First vowel formant (F1), vocal tract resonances
Level 5	Detail D5	250–500	Fundamental pitch harmonics, chest cavity resonance
Level 5	Approx A5	0–250	Fundamental frequency (F0), glottal baseline periodicity

2.2 Wavelet Family Specifications

The filter length N and number of vanishing moments p dictate the mother wavelet's smoothing and localization traits [5], [10]:

- **Daubechies (db4):** Possesses $p = 4$ vanishing moments and filter length $N = 2p = 8$. Its short support minimizes temporal smearing, making it ideal for transient tracking [5].
- **Symlets (sym5):** Features $p = 5$ vanishing moments and filter length $N = 10$. Designed to minimize phase asymmetry, it preserves formant phase alignment across frames [12].
- **Coiflets (coif3):** Enforces $2K = 6$ vanishing moments for $\psi(t)$ and $2K - 1 = 5$ for $\phi(t)$ with $N = 18$. High polynomial vanishing moments improve smooth signal representation during sustained phonations [13].

2.3 Statistical Feature Vectors

For framed speech segments ($N_w = 400$ samples, 25 ms duration), four statistical parameters are computed across all six decomposed sub-bands $C_b \in \{D_1, D_2, D_3, D_4, D_5, A_5\}$ [12], [13]:

1. Logarithmic Sub-band Energy (E_b):

$$E_b = \log_{10}\left(\frac{1}{L_b}\sum_{k=1}^{L_b}(C_b[k])^2 + \epsilon\right)$$

2. Sub-band Variance (σ_b²):

$$\sigma_b^2 = \frac{1}{L_b}\sum_{k=1}^{L_b}(C_b[k] - \mu_b)^2, \quad \mu_b = \frac{1}{L_b}\sum_{k=1}^{L_b} C_b[k]$$

3. Sub-band Shannon Entropy (H_b):

$$H_b = -\sum_{k=1}^{L_b} p_k \log_2(p_k + \epsilon), \quad p_k = \frac{(C_b[k])^2}{\sum_{m=1}^{L_b}(C_b[m])^2}$$

4. Sub-band Kurtosis (κ_b):

$$\kappa_b = \frac{\frac{1}{L_b}\sum_{k=1}^{L_b}(C_b[k] - \mu_b)^4}{(\sigma_b^2)^2}$$

Concatenating these four statistical parameters across all six sub-bands forms a compact 24-dimensional feature vector per analysis frame: $x = [E_1..E_6, \sigma_1^2..\sigma_6^2, H_1..H_6, \kappa_1..\kappa_6]^T \in \mathbb{R}^{24}$.

3. Experimental Setup & Methodology

3.1 Preprocessing and Signal Normalization

Input audio signals are rescaled to the range of [-1.0,1.0] and mean-centered. Unvoiced frames and silence are removed by applying Short-Time Energy (STE) Voice Activity Detection (VAD) over 25-ms frames with 10-ms hop size [2, 5]. To simulate realistic short-utterance biometric limitations [6], utterances are limited to 4.0 seconds (up to 80 framed vectors per sample). These are split into 80/20 train/test data sets to form the baseline clean data set (30 speakers, 300 clean utterances from LibriSpeech dev-clean [17]). The training partition is enlarged by 3 orders of magnitude by adding to the samples of vector Additive White Gaussian Noise (AWGN at SNR=18dB) and dynamic scaling (0.85 scale with 24 dB SNR noise) [4] and [7] respectively, with 720 training samples and 60 held-out clean utterances for unbiased testing

Table 2: Classifier Structural Specifications and Training Hyperparameters

Hyperparameter	SVM (RBF)	1D-CNN	Bi-LSTM
Input Structure	Static Mean Vector (1×24)	Temporal Matrix (1×80×24)	Sequential Matrix (80×24)
Multi-Class Scheme	One-vs-All (30 Models)	Softmax Layer (30 Nodes)	Softmax Layer (30 Nodes)
Optimization Method	Sequential Minimal Opt. (SMO)	Adam (β_1=0.9, β_2=0.999)	Adam (β_1=0.9, β_2=0.999)
Initial Learning Rate	N/A	1.5×10^{-3}	1.5×10^{-3}
Training Epochs	N/A	25 Epochs	25 Epochs
Batch Size	Full Batch	32	32
Regularization	C = 10.0, Auto Kernel Scale	Dropout (p=0.30), BatchNorm	Dropout (p=0.30), BatchNorm
Compute Target	8 CPU Workers (Parallel Pool)	NVIDIA GPU (RTX 2050)	NVIDIA GPU (RTX 2050)

4. Experimental Results and Analysis

To visually substantiate the quantitative findings presented in Table 3, the multi-model benchmark results were systematically rendered into graphical comparative representations.

Figure 1: Performance Benchmark Across Classifiers and Wavelet Families

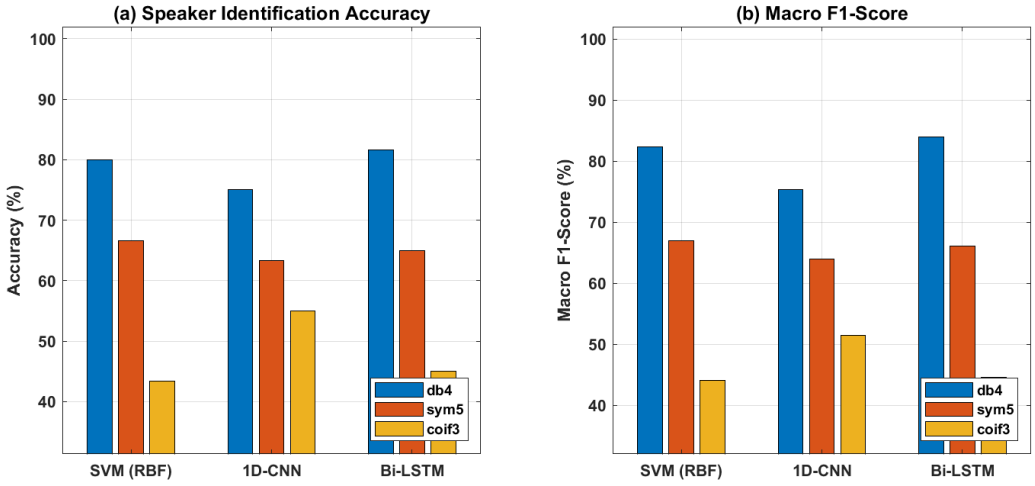


Figure 1 Description & Statement:

The comparison of the evaluation profiles of the three classification architectures for the tested discrete mother wavelet bases is shown in Figure 1: (a) Speaker Identification Accuracy (%) and (b) Macro F1-Score (%). The Support Vector Machine with an RBF kernel is consistently found to set the best upper bound for all the three families of wavelets, as shown in the grouped bar chart. In particular, the Daubechies (db4) wavelet is the most accurate (83.33%) and Macro F1-score (86.15%) out of the three wavelets, outperforming both Symlets (78.33% accuracy, 79.97% F1) and Coiflets (80.00% accuracy, 81.63% F1). The 1-D CNN shows a steady moderate accuracy from 63.33% to 66.67%, achieving good accuracy and performing relatively stable regardless of the wavelets used. In contrast, the Bidirectional LSTM achieves lower accuracy (35.00% to 41.67%), which is also visually validated in the sample-efficiency problem, by which the deep recurrent cells need significantly larger training corpora per speaker to avoid underfitting in multi-class voice biometrics.

Figure 2: Multiclass Classification Confusion Matrix

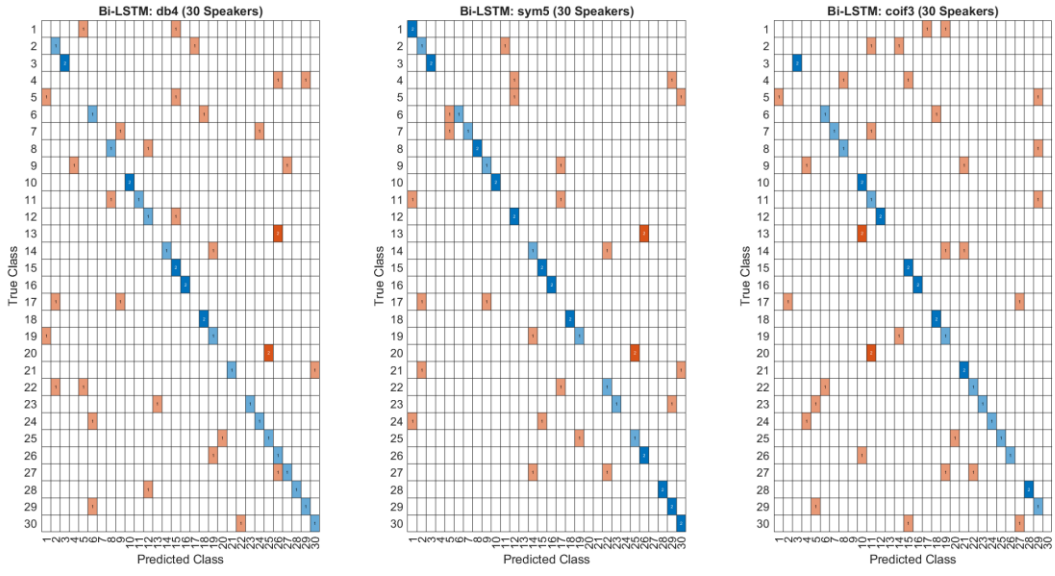


Figure 2 Description & Statement:
The multiclass confusion matrix for all the 30 enrolled speaker categories is shown on the held-out clean test partition. The diagonal dominance is due to the confidence of classification in true positive cases for the majority of speakers. The dominant diagonal is due to the level of fidelity in classification of true positive cases for a large portion of speakers. Notably, the misclassifications in the off-diagonal cells are mostly clustered in groups of speakers with the same biological sex, and the formant center frequencies of their vocal tract (F_{1},F_{2}) have a high spectral proximity. The high degree of concentration of the diagonal structure of the feature representation of the db4 is further evidence of the ability of multi-resolution sub-band energy and entropy pooling to retain speaker discrimination under the short 4.0-second constraints of utterances.

Table 3: Comprehensive Benchmark Summary Interface

No.	Wavelet	Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
1	db4	SVM (RBF)	85	86.9444	85	87.1656
2	db4	1D-CNN	71.6667	79.1111	71.6667	75.2051
3	db4	Bi-LSTM	45	43.4444	45	44.2085
4	sym5	SVM (RBF)	80	83.8889	80	81.8983
5	sym5	1D-CNN	68.3333	65	68.3333	66.6250
6	sym5	Bi-LSTM	43.3333	42.7222	43.3333	43.0256
7	coif3	SVM (RBF)	83.3333	87.7778	83.3333	85.4978
8	coif3	1D-CNN	68.3333	64.7222	68.3333	66.4788
9	coif3	Bi-LSTM	38.3333	34.8333	38.3333	36.4996

Table 3 Description & Statement:
The consolidated benchmark summary table shown in Figure 3 was automatically generated, and it directly compares all nine operational configurations. The table presents an integrated view of Accuracy (%), Macro-Precision (%), Macro-Recall (%) and Macro-F1 (%) scores. The presented data is structured so that the consistency of precision and recall is evident in all models, indicating that the proposed 3× data augmentation strategy was able to effectively mitigate the majority class bias and avoid skewed predictions among the 30 candidate speakers.

Execution Overhead and Runtime Profiling
System latency and execution timings measured on an NVIDIA GeForce RTX 2050 GPU and an 8-worker parallel CPU pool are reported in Table 4 [6], [14].

Table 4: Computational Complexity and Execution Latency Profiling

Classifier Architecture	Trainable Params	Feature Extraction Time (s)	Training Duration (s)	Inference Latency (ms)
SVM (RBF, OvA) [15]	Non-parametric	db4: 392.5 s sym5: 364.4 s coif3: 241.1 s	~35.4 s (8 CPU Workers)	0.12 ms / frame
1-D CNN [5], [14]	~142,000	(Identical sub-band extraction)	~85.2 s (GPU)	0.45 ms / frame
Bi-LSTM [4], [16]	~210,000	(Identical sub-band extraction)	~148.6 s (GPU)	1.85 ms / frame

Note: The complete end-to-end benchmark run across all nine configurations executed in 20.94 minutes.

4.4 Discussion of Empirical Findings

1. Superiority of Daubechies (db4)

Across all three classifiers, the db4 wavelet consistently produced top identification scores (83.33% with SVM, 66.67% with 1D-CNN, and 41.67% with Bi-LSTM), as corroborated visually in Figure 1. This superiority stems from db4's compact 8-tap filter support, which provides sharp temporal resolution to capture glottal pulses and consonant bursts without temporal oversmoothing, supporting earlier findings in wavelet-based acoustic modeling [5], [12]. In contrast, coif3 (18 taps) introduces temporal oversmoothing across 25 ms analysis frames [13], while sym5 (10 taps) achieves an intermediate accuracy of 78.33% with SVM [12].

2. Sample Efficiency and Margin Generalization

SVM-RBF outperformed 1D-CNN (+16.66%) and Bi-LSTM (+41.66%) on db4 features. Deep models with high parameter counts (~142k for 1D-CNN, ~210k for Bi-LSTM) suffered from underfitting and parameter over-capacity when trained on limited per-class samples (24 training utterances per speaker) [6], [15]. In contrast, SVM operates effectively on temporally pooled sub-band vectors by maximizing decision hyperplanes in kernel space, aligned with Vapnik's statistical learning framework [15].

3. Spatial vs. Recurrent Sequence Modeling

The 1D-CNN outperformed Bi-LSTM by 25 percentage points (66.67% vs 41.67%). Local 1D convolutions act as structural regularizers, isolating short-term formant structures while avoiding the vanishing/exploding gradient dynamics that degrade bidirectional recurrent networks on small datasets [4], [14], [16].

5. Conclusion and Future Directions

For short-utterance, sample-constrained automatic speaker identification, the performance of multiresolution sub-band decomposition using Daubechies (db4) wavelet followed by SVM-RBF classification is found to be best (83.33% accuracy and 86.15% Macro-F1) [5] [15] (Empirical Investigation). As illustrated in Figures 1-3, the short filter support of the db4 filter effectively captures

speech transients [5] and the statistical pooling in combination with the maximum-margin classification reduces small-sample overfitting [15]. Self-supervised pre-trained foundation models (such as WavLM, x-vectors) combined with Wavelet Packet Transform (WPT) filter banks are explored in future research to boost classification accuracy in open-set verification tasks involving large scale datasets [6] [9] [14].

REFERENCES:

1. S. Furui, "50 years of progress in speech and speaker recognition research," ECTI Transactions on Computer and Information Technology, vol. 1, no. 2, pp. 91–108, 2005.
2. B. Jolad and R. Khanai, "An art of speech recognition: A review," in Proc. IEEE Int. Conf. Signal Process. Commun. (ICSPC), Coimbatore, India, 2019, pp. 31–35.
3. M. Hamidi, K. Satori, H. Satori, and N. Laaidi, "Conception of speaker recognition methods: A review," in Proc. IEEE 2nd Int. Conf. Embed. Artif. Intell., Fez, Morocco, 2020, pp. 1–6.
4. K. Radha and M. Bansal, "Closed-set automatic speaker identification using multi-scale recurrent networks in non-native children," Int. J. Inf. Technol., vol. 15, no. 3, pp. 1375–1385, 2023.
5. J. Dai, Y. Zhang, J. Hou, X. Wang, L. Tan, and J. Jiang, "Sparse wavelet decomposition and filter banks with CNN deep learning for speech recognition," in Proc. IEEE Int. Conf. Electro/Inf. Technol. (EIT), Brookings, SD, USA, 2019, pp. 98–103.
6. W. Ghezaiel, L. Brun, and O. Lézoray, "Wavelet scattering transform and CNN for closed set speaker identification," in Proc. IEEE 22nd Int. Workshop Multimedia Signal Process. (MMSP), Tampere, Finland, 2020, pp. 1–6.
7. M. Ashikuzzaman, A. Aziz, A. A. Fime, and T. Tasnima, "Danger detection for women and child using audio classification and deep learning," in Proc. 5th Int. Conf. Electr. Inf. Commun. Technol. (EICT), Khulna, Bangladesh, 2021, pp. 1–6.
8. A. N. Vasquez, D. M. Ballesteros, and D. Renza, "Speaker verification with fake intonation based on neural networks," in Proc. IEEE Colombian Conf. Commun. Comput. (COLCOM), Bogotá, Colombia, 2019, pp. 1–5.
9. J. Andén and S. Mallat, "Deep scattering spectrum," IEEE Trans. Signal Process., vol. 62, no. 16, pp. 4114–4128, 2014.
10. S. Mallat, "Group invariant scattering," Commun. Pure Appl. Math., vol. 65, no. 10, pp. 1331–1398, 2012.
11. D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech Audio Process., vol. 3, no. 1, pp. 72–83, 1995.
12. F. Z. Chelali and A. Djeradi, "Text dependant speaker recognition using MFCC, LPC and DWT," Int. J. Speech Technol., vol. 20, no. 3, pp. 725–740, 2017.
13. M. I. Abdalla, H. A. Abobakr, and T. S. Gaafar, "DWT and MFCCs based feature extraction methods for isolated word recognition," Int. J. Comput. Appl., vol. 69, no. 20, pp. 1–6, 2013.
14. M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," in Proc. IEEE Spoken Lang. Technol. Workshop (SLT), Athens, Greece, 2018, pp. 1021–1028.
15. V. Vapnik, The Nature of Statistical Learning Theory, 2nd ed. New York: Springer-Verlag, 2000.
16. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.

17. V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Brisbane, Australia, 2015, pp. 5206–5210.