Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.036797
Article

An Efficient Text-Independent Speaker Identification Using Feature Fusion and Transformer Model

Arfat Ahmad Khan[1], Rashid Jahangir[2,*], Roobaea Alroobaea[3], Saleh Yahya Alyahyan[4], Ahmed H. Almulhi[3], Majed Alsafyani[3] and Chitapong Wechtaisong[5,*]

[1]College of Computing, Khon Kaen University, Khon Kaen, 40000, Thailand
[2]Department of Computer Science, COMSATS University Islamabad, Vehari Campus, Vehari, 61100, Pakistan
[3]Department of Computer Science, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif, 21944, Saudi Arabia
[4]Department of Computer Science, Community College in Dwadmi, Sharqa University, Dawadmi, 17472, Saudi Arabia
[5]School of Telecommunication Engineering, Suranaree University of Technology, Nakhon Ratchasima, 30000, Thailand
*Corresponding Authors: Rashid Jahangir. Email: rashidjahangir@cuivehari.edu.pk; Chitapong Wechtaisong. Email: chitapong@g.sut.ac.th

Received: 12 October 2022; Accepted: 30 January 2023

Abstract: Automatic Speaker Identification (ASI) involves the process of distinguishing an audio stream associated with numerous speakers' utterances. Some common aspects, such as the framework difference, overlapping of different sound events, and the presence of various sound sources during recording, make the ASI task much more complicated and complex. This research proposes a deep learning model to improve the accuracy of the ASI system and reduce the model training time under limited computation resources. In this research, the performance of the transformer model is investigated. Seven audio features, chromagram, Mel-spectrogram, tonnetz, Mel-Frequency Cepstral Coefficients (MFCCs), delta MFCCs, delta-delta MFCCs and spectral contrast, are extracted from the ELSDSR, CSTR-VCTK, and Ar-DAD, datasets. The evaluation of various experiments demonstrates that the best performance was achieved by the proposed transformer model using seven audio features on all datasets. For ELSDSR, CSTR-VCTK, and Ar-DAD, the highest attained accuracies are 0.99, 0.97, and 0.99, respectively. The experimental results reveal that the proposed technique can achieve the best performance for ASI problems.

Keywords: Speaker identification; signal processing; Arabic; deep learning; transformer

1 Introduction

Although, there are several ways available to exchange information among people, like text messages and emails. In recent years some visual methods have also been utilized to show expressions, such as pictures, emojis, and stickers, while communicating with each other. But these all are supplementary. Yet, speech remains one of the most effective ways of communication among humans due to its exclusive data-enriched properties. Both linguistic and para lingual features can be extracted from speech, which makes it distinct from other communication methods. In recent decades, scientists have dedicated considerable attention to studying human voice and speech. Since this study focuses on the speech signal, it discusses speech signal production and its perception in the following sections.

1.1 Speech Production

Fig. 1 shows the human speech/voice production system. This system can produce multiple sounds. Speech production, however, mainly depends on three subsystems: the laryngeal subsystem (larynx), the respiratory subsystem (lungs, diaphragm), and the articulatory subsystem (hard/soft palate, nasal/oral cavities, jaw, tongue, teeth, and lips) [1]. The speech signal production formation begins when air moves upward from the lungs into the larynx. Next, it passes through the trachea, pharynx, and oral and nasal cavities. Finally, after cavities, it passes through the lips to generate vocal sounds [2].

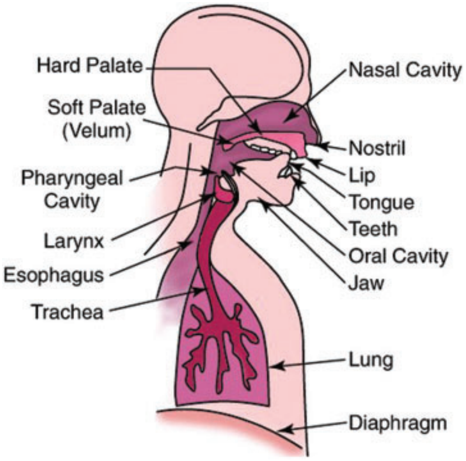


Figure 1: An overview of the human voice production system

The generation of sound depends on the vibration of vocal cords, either voiced or unvoiced [3]. The fluctuation of vocal cords depends on the tightness of cords that the muscles [in folds] and the mass of cords control. However, the Bernoulli effect of air also affects the fluctuations. Later, airstream impulses are broken by cords' opening and closing. The shape and cycle of the impulses are determined by the pitch and loudness of the speech signal. Thus, the pitch of the speech signal is the fundamental frequency of the glottal pulse.

1.2 Speech Perception

Scientists use clues of human speech perception mechanisms for feature-level representation. Fig. 2 provides basic information about a human's ear's interior settings that comprise three parts: outer, middle, and inner.

The outer part—auricle or pinna—is connected to a short exogenous hearing canal; a membrane, the eardrum, closes its end. The outer part is the gateway of sound that travels through the ear canal to reach the middle part [4]. Once the sound waves reach the eardrum, the latter vibrates. Then the waves hit three small bones—stapes (stirrup), incus (anvil), and malleus (hammer)—which are in the middle part. The main task of the middle part is to turn sound waves into vibrations and send them to the inner part of the ear. Thus, in addition to delivering sounds, the middle part shields human hearing from high-intensity sounds [5].

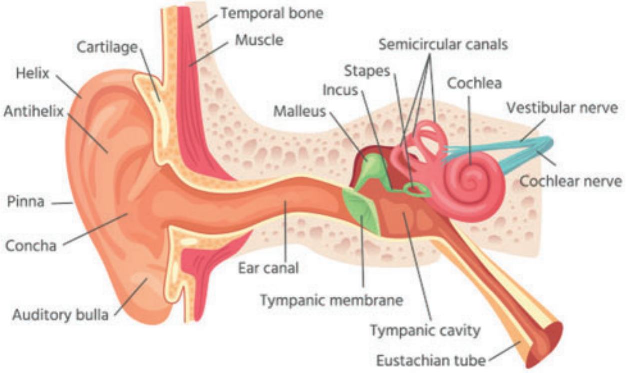


Figure 2: Human ear anatomy

A spiral-shaped cavity, cochlea, is the primary component of the [liquid-filled] inner part. The cochlea is padded with thousands of microscopic hairy cells. After the sound vibrations reach liquid, the latter, contingent on the inbound sound, vibrates in multiple patterns. The ensuing vibration causes movement in the cochlea hair that transforms vibration into a nerve signal. Then through the hearing nerve, the nerve signal is delivered to the brain.

Security of information is one of the vital areas of research these days. For security, human biometrics are usually used as they distinguish between individuals. In the context of biometrics, an automated recognition of different human biological and behavioral identifications is made, which may include human eyes (cornea, iris), human gestures, fingerprints, face, voice, and Deoxyribonucleic Acid (DNA) analysis. The human voice is one of the communicative biometrics that holds information about a person's distinctive traits, such as identity, age, gender, and emotion. Voice, as a biometric for identifying a particular human, is known as speaker identification in academic literature. Speaker identification can be applied in various applications while being an accurate source of human identification, a suitable and convenient technology. It is more likely to apply speaker identification to several applications to authenticate humans. Some suitable applications might include forensic voice verification to detect suspects by government law enforcement agencies [6,7], access control to different services, such as telephone network services [8], voice dialing, computer access control [9], mobile banking, and mobile shopping. Moreover, speaker identification systems are extensively used to improve security [10], automatic speaker labeling of recorded meetings, and personalized caller identification using intelligent answering machines [11]. These applications are a source of motivation to design and develop an automated system for speaker identification.

Voice recognition system (VRS) in mobile has better market advantages than same system installed on a PC. A VRS in smartphones, or other 3C portable devices can process the user's voice and understand its commands. With such a recognition system, some practical algorithms can be applied to devices mentioned above, not only to protect the devices from unauthorized use but also to make them convenient and friendly. Generally, a Personal Computer (PC) has better computing power and capabilities than a mobile device. So, running the algorithms of voice recognition on a mobile device can be a difficult task. To imply these algorithms in a mobile device, the mobile needs better operational efficiency, processing capability, and the number of resources it consumes [12].

In this study, acoustic features are extracted from each audio file for automatic speaker identification. These features include spectral contrast, Mel-scaled spectrogram, tonnetz representations, chromagram, Mel-frequency cepstral coefficients (MFCCs), delta (Δ) MFCCs, and delta-delta ($\Delta\Delta$) MFCCs. Afterward, the derived features were fed as input to the transformer model. To evaluate the performance of the transformer model, this study employed three datasets: the English Language Speech Database for Speaker Recognition (ELSDSR), the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), and the Surrey Audio-Visual Expressed Emotion (SAVEE). The proposed method yields better identification performance for an automatic speaker identification system. In this work, an efficient algorithm to derive discriminative and salient features for automatic speaker identification is proposed. Moreover, the proposed transformed model achieved better speaker identification performance over the existing baseline methods.

The rest of this work is organized as follows: Section 2 presents the literature review on automatic speaker identification methods. Section 3 describes the datasets utilized to perform experiments, the feature retrieval process, and deep learning-based transformer model adopted for automatic speaker identification and evaluation parameters. The results of different experiments and the significance of the observed findings is presented in Section 4. Finally, Section 5 concludes this study.

2 Literature Review

In literature, automatically identifying speakers can be divided into two stages. To identify the speaker from speech signals, the first step is to extract valuable and discriminative features from the speaker's utterances. The second step is to select the classification algorithm. Below is a quick recap of these two steps for automatic speaker recognition.

2.1 Feature Extraction

Two approaches have been widely employed in existing studies to extract the feature from audio signals. The most commonly utilized approach for feature extraction is to split the Audio signal into various frames of a specific length and retrieve low-local features from all frames. Generally, the features used for automatic speaker identification are categorized into four types: linguistic feature, acoustic feature, contextual feature, and hybrid feature. Acoustic properties are the best and most widely used for automatic speaker identification. They consist of spectral features, voice quality features, and prosodic features (duration, loudness, and pitch) (MFCCs etc.). In [13] the authors have presented the fusion of MFCCs and 12 time-domain features (MFCCT). Using the LibriSpeech dataset, the MFCCT feature was provided as input to a deep neural network in the second phase, which resulted in 93% accuracy.

Due to the limits of machine learning classifiers to handle massive databases and considerable advancements in computer capability, deep learning techniques for automatic speaker recognition have been gaining attention recently. To do this, [13] proposed a gated recurrent unit (GRU) and 2D Convolutional Neural Network (CNN) for automatic speaker identification. In the pro-posed model, the convolutional layer was used to extract the voiceprint features and reduce the dimensionality in both frequency and time domains. The GRU layer was used to fast the computation process. The authors of this study evaluated various network structures including, deep Recurrent Neural Network (RNN), 2D CNN and Long Short-Term Memory (LSTM) on the Aishell-1 database. The experimental results revealed that the proposed GRU model obtained an accuracy of 98.9%. In another study, [14] conducted a thorough analysis of deep learning techniques such as CNN, RNN, Deep Belief Network (DBN), and Restricted Boltzmann Machine (RBM) with their features, advantages, and disadvantages. Also covered was a survey of several AI algorithms for speaker recognition on demand. In another article, Bai et al. [15] reviewed different subtasks of automatic speaker identification and focused on deep learning approaches. The authors discussed the significant advantages of deep learning approaches over classical machine learning methods. The deep learning approaches can extract highly abstract features from speaker utterances. The authors of this study paid close attention to fundamental components of speaker identification involving the inputs, structures of the network for feature extraction, pooling strategies, and impartial functions, respectively. Moreover, this survey also focused on speech enhancement and domain adaptation to deal with noise issues and domain mismatch. Another study [16] presented a novel method to enhance the performance of an automatic speaker identification systems in the presence of interference using CNN for robot applications. Firstly, the authors divided the audio signal into frames, each of which was converted into a spectrogram and consequently, Radon transformation was estimated for this spectrogram. Afterward, spectrograms and their Radon transformation were utilized as input to the CNN model. The proposed CNN model comprised of six convolutional layers (CLs) followed by six Max. Pooling layers. The experimental results revealed that the proposed model achieved a high accuracy of up to 97%.

An et al. [17] presented two variations of CNN for automatic feature learning, the residual neural networks (ResNets) and the visual geometry group (VGG) nets using a self-attention layer. The weighted average over all time steps was learned by this layer. In addition to handling varied length frames, the suggested model with self-attention structure also acquired speaker characteristics from diverse angles. The proposed model was evaluated using the benchmark dataset, VoxCeleb and experimental results showed the superiority over the traditional i-vector speaker identification models. In another study [18], authors conducted experiments using little training data and a deep learning approach. In this approach, authors omitted the pre-processing step and considered large segments of speech signals. The proposed Bidirectional Long Short-Term Memory (BLSTM) achieved an accuracy of 77% on individual speech segments and 99.5% when segments of each speaker were considered as a bundle. To identify the speakers in stressful environments, Nassif et al. [19] presented an effective technique called radial basis function neural network-CNN. In this study, the proposed model was evaluated on Arabic Emirati-accent and English Speech Under Simulated and Actual Stress (SUSAS) databases. The authors extracted the MFCCs feature and fed it as input to the proposed model. Comparing the proposed model and traditional machine learning classifiers revealed that the proposed model outperformed the other methods in stressful environments. Another study Nassif et al. [20], evaluated the performance of automatic speaker identification in real-time situations such as emotional and noisy conditions. The authors incorporated two modules: a pre-processing module called computational auditory scene analysis and a cascaded Gaussian mixture model (GMM)-CNN model for ASI. The experimental results demonstrated that the proposed method produced promising results and outperformed other classifiers.

2.2 Classification Methods

Multiple machine learning-based classifiers, including the GMM, hidden Markov model (HMM) [21], multilayer perceptron (MLP), k-nearest neighbor (k-NN) [22], support vector machine (SVM) [23], and random forest (RF), have been used by many researchers to identify speakers from audio data signals. These classifiers have been extensively used in speech-related applications, including automatic speaker identification and emotion recognition. The classifiers RF, J48, k-NN, Naive Bayes (NB), and SVM are employed in this work. These classifiers' performance is assessed for accuracy and contrasted with the performance of the transformer model.

3 Proposed Methodology

This section presents the general research methodology for developing the features fusion and transformer model for the ASI system. Recently, various techniques have been proposed for developing the ASI system. However, these techniques present some drawbacks that have hindered their efficient implementation. Therefore, two techniques identified to reduce the limitations inherent in the current systems include feature fusion and transformer model for automatic feature representation. Feature fusion method handles high dimensional data, increases reliability, and improves the generalization of the ASI system. Conversely, the transformer model is an automated feature representation method that uses hierarchical layers to extract the discriminative features from speech data. The proposed research methodology consists of five different processes. These include the methodology employed for the research, data collection, feature extraction, brief descriptions of the proposed model, analysis, and performance evaluation techniques. The high-level description of the proposed ASI method is presented in Fig. 3. All experiments are performed on a laptop with 64-bit Windows 11 OS, 8 GB RAM, Intel® Core (TM) i5-3210M CPU, and Spyder (Python 3.8.5) environment. A detailed explanation of the proposed method is given in subsequent subsections.

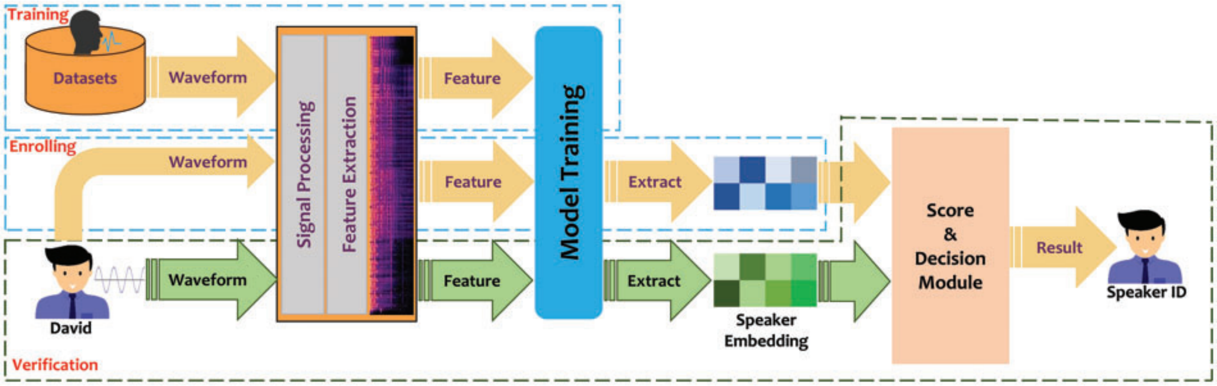


Figure 3: Proposed research methodology for ASI

3.1 Datasets Description

The benchmark datasets were utilized for experiments conducted in this research. The speech datasets were collected from several male and female speakers under different conditions, recording devices, technical settings, number of sessions, demography, and linguistic variability. The characteristic of each dataset is explained in detail below. The datasets utilized in this research provide various characteristics to ensure the comprehensive implementation of the proposed techniques. Unlike datasets employed by recent studies [24,25] for deep learning-based speaker identification, this study comprehensively analyses three datasets of different conditions.

ELSDSR is a speech dataset in the English language that was created for the use and testing of SI and accent recognition systems. 10 female students and 12 male students and researchers at the Technical University of Denmark's chamber building participated in a single recording session (DTU). Among a total of 22 speakers, 20 were Dane, one was an Icelander, and the other was a Canadian. The average age of male speakers was lower than that of a female because of unequal gender distribution at the recording site. The utterances were captured using the Pulse Coded Modulation (PCM) method at a sampling rate of 16 kHz and a bit rate of 16 bits on a MARANTZ PMD670 recorder into the most common .wav file format. The corpus is then separated into 44 (2 × 22) test utterances and 154

(7 × 22) training utterances. The average time for reading the training data is 88.3 s for females, 78.6 s for males, 83 s for all speakers. Similarly, the average duration for test data is 19.6 s (female), 16.1 s (male), 17.6 s (for all speakers). Table 1 shows the reading duration spent on both training and test data individually.

Table 1: Reading duration for training and test data

No	Female	Train (s)	Test (s)	Male	Train (s)	Test (s)
1	FAML	99.1	18.7	MASM	81.2	20.9
2	FHRO	86.6	21.2	MFKC	91.6	15.8
3	FDHH	77.3	12.7	MCBR	68.4	13.1
4	FEAB	92.8	24.0	MLKH	76.8	14.7
5	FMEL	76.3	18.2	MKBP	69.9	15.8
6	FJAZ	79.2	18.0	MMNA	73.1	10.9
7	FSLJ	80.2	18.4	MMLP	79.6	13.3
8	FMEV	99.1	24.1	MOEW	88.0	23.4
9	FUAN	89.5	25.1	MNHP	82.9	20.3
10	FTEJ	102.9	15.8	MREM	79.1	21.8
11				MPRA	86.8	9.3
12				MTLS	66.2	14.05

CSTR-VCTK (Centre for Speech Technology Voice Cloning Toolkit) is a dataset that holds the speech data of 109 native English speakers having different accents. Every speaker reads almost 400 sentences, including a passage from The Herald (Glasgow) newspaper and the Rainbow Passage. In addition, to distinguish the speaker's accent, it included an elicitation paragraph. The selected text from the newspaper was granted permission from Herald & Times Group. Moreover, every speaker recorded a different set of sentences compared to the other speakers. For the selection of the set of sentences and to maximize contextual and phonetic coverage, a greedy algorithm was employed. Conversely, all 109 speakers used the same elicitation paragraph and Rainbow Passage in their recorded utterances.

Ar-DAD (Arabic diversified audio dataset): The audio clips include two subdirectories named reciters and imitators. Besides, it has a third directory, including similar Quran verses as plain text. The reciters' directory contains 37 folders and 15,810 files. All files in the reciter's directory are arranged as chapters, and each chapter contains verses, and each verse includes reciters that cover Quranic chapters starting from 74 to 114. Each chapter has subfolders of verses, and all verse folders contain 30 audio clips. The data format is in WAV form with a 44.1 kHz sampling rate, stereo channel, and 16-bit depth. Both manual and acoustic approaches were used to create the WAV format because it is acceptable by well-known machine learning algorithms.

3.2 Feature Extraction

At this phase, features are extracted from the collected dataset. The extraction of salient and discriminative features that correctly identify speakers from voice is the most critical phase in the success of the automatic speaker identification system [26]. The selection of suitable features can significantly enhance the performance of the ASI system, while irrelevant features can delay the training process of the model [27]. In this research, Librosa [28] audio library was used to retrieve the feature representations of an audio signals. These representations involve:

- Spectral contrast
- Mel-spectrogram
- Tonnetz representation
- Chromagram
- MFFCs
- ΔMFCCs
- ΔΔMFCCs

The most common applications for MFCCs features are automatic speaker identification and voice recognition [29]. A speech signal is initially broken into units of significance called frames to recover the MFCC's characteristics. Second, the stillness at the beginning and conclusion of each frame is lessened by using a widely used technique called windowing operation. The Fast Fourier Transform is then used to transform these frames from the time-domain to the frequency-domain (FFT). Using Eq. (1), the frequency values generated by the FFT are evaluated on the Mel-scale. After calculating the logs of the powers at each Mel-frequency, the Discrete Cosine Transform is used to convert each log Mel-spectrum into a time-domain (DCT). The MFCCs are the computed amplitudes from the resulting spectrum. Although MFCCs features are helpful for detecting and tracking timbre variations in voice signals, they have trouble differentiating between pitches and harmony classes. Chroma features are extracted from the speech stream using binning techniques and short-time Fourier transform to solve this issue. All chromagram features for each frame are recorded and converted to one coefficient in this work. An audio file is first separated into frames to obtain the Mel-spectrogram feature, and then FFT is computed for each frame. After that, a Mel-scale is created by splitting the frequency spectrum into equally spaced frequencies.

Finally, the frequencies are calculated on a Mel-scale for each audio frame. The Tonnetz representation, which displays the pitch relations in the rise and fall of voice signals, is a 6-dimensional tonal centroid represented by the Harmonic Network. For each frame of the speech signal, the pitch space (tonal centroid) features are computed in this work. By computing the root mean square divergence between the spectral peak and the spectral depression for each frame, spectral contrast produces an inclusive spectral proof of speech signal. To combine the many voice qualities, such as pitch, harmony, timber, etc., into a single training utterance, 273 features (7 spectral contrast, 128 Mel-spectrogram, 6 tonnetz, 12 chromagram, 40 MFCCs, 40 delta MFCCs, 40 delta-delta MFCCs) are retrieved in this work. Algorithm 1 represents the code to compute the master feature vector for input to the transformer model.

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

3.3 Transformer Model

When learning long sequences, vanishing gradient is a prevalent problem in Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) models. Though LSTM has addressed the issue of RNN by using the carry-forward technique, which contains the information of previous hidden layers and previous of the previous hidden layers. Nonetheless, when it comes to longer sequential issues, this LSTM carrying forward technique may not work. Without employing RNNs or aligned convolution, the transformer model with a self-attention mechanism is used to calculate the representations of input and output. The technique of a self-attention to calculate the representation of a sequence, the self-attention mechanism links to various points of that sequence. Transformers have stack-based encoder-decoder modules with a self-attention mechanism as its architecture.

Encoder: The input sequence is converted to a continuous representation sequence using the encoder. Every time a step is produced, the transformer adds the previously generated representation as an extra input sequence. This work's encoder consisted of a stack of six identical layers, each was further subdivided into two layers: a multi-head self-attention layer and a fully connected feed-forward network. For every two sub-layers, a residual connection was used, followed by normalization.

Decoder: A stack of 6 similar layers made up the decoder as well. To guarantee that any sequence predictions are based only on tokens that occurred before the current token, a third sub-layer known as masked multi-headed attention was added to the decoder. Finally, residual connections were used around each sub-layer, followed by the normalization layer.

Attention: It allows the transformer model to concentrate on other input audio sequences related to that word. The self-attention is efficient in maintaining the context-based feature in an audio sequence. These features are derived using a set of Quires Q, keys K, and values as given in Eq. (2):

$$Attention\ (Q,\ K,\ V) = Softmax\ (QKT)\ V \tag{2}$$

A matrix containing the details of each audio sequence is the result of a given equation. Transformers can parallelize training since, Q, K, and V are stacked as a matrix. A fully connected feed-forward network was also applied to every point individually and uniformly in every encoder and decoder layer. Here, a ReLU activation separates two linear transformations. The input and output tokens were converted to dimensional vectors by using learned embeddings as well. To calculate the class probabilities, the decoder output was transformed using a softmax function. The general structure of the transformer model is shown in Fig. 4.

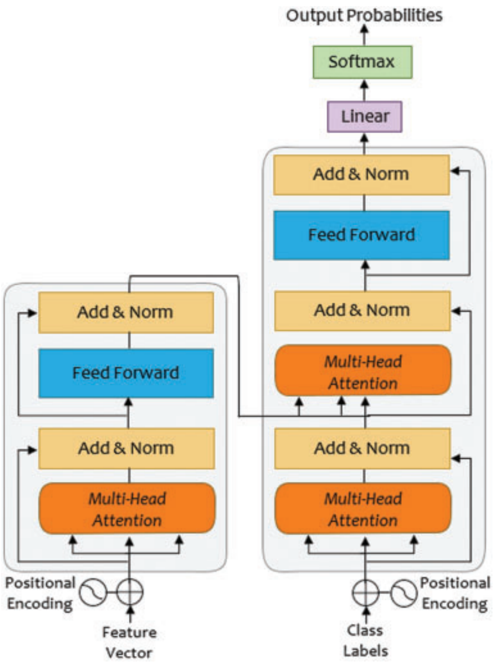


Figure 4: Structure of transformer model

Algorithm 1: Algorithm to Compute LIST of Features

```
Input: Path to folder of speaker dataset
Output: list of features for the input to transformer
N ← number of speakers in path dataset
FOR i := 1 to N DO LOOP
    total ← number of audio files of single speaker
    FOR j := 1 to total DO LOOP
        X ← load audio file j.
        SC ← spectral contrast from X
        Mel ← melspectrogram from X
        Tonnetz ← tonnetz representations from X
        chroma ← chromagram from X
        mfcc ← Mel-Frequency Cepstral Coefficients (MFCCs) from X
        delta ← $\Delta$MFCCs from X
        delta-delta ← $\Delta\Delta$MFCCs from X
        features ← hstack(SC, Mel, Tonnetz, chroma, mfcc, delta, delta-delta)
        arr ← features, i
        list ← arr
        j ← j + 1
    END
    i ← i + 1
END
```

3.4 Evaluation Metrics

This study used accuracy, recall, precision, and F1-score as its four main assessment metrics to assess the effectiveness of the automatic speaker identification model. The confusion matrix, which is a table of false-positive (FP), false-negative (FN), and true-positive (TP), true-negative (TN), when the classification algorithm effectively detects, was used to calculate the performance of all classes (Speaker ID). Accuracy uses the total number of utterances to calculate the accurately identified classes (Speaker ID) using Eq. (3).

$$Accuracy = \frac{1}{N}\sum\nolimits_{i=1}^{N} \frac{(TP + TN)_i}{(TP + TN + FP + FN)_i} \tag{3}$$

where N represents the number of utterances.

According to Eq. (4), the fraction of properly identified positive instances and the sum of correctly and mistakenly identified positive and negative instances are used to compute a class's recall. It stands for the speaker identification model's completion.

$$Recall = \frac{1}{N}\sum\nolimits_{i=1}^{N} \frac{(TP)_i}{(TP + FN)_i} \tag{4}$$

Precision of a class is calculated as the fraction of correctly identified positive instances and the total number of correctly and incorrectly identified positive instances as given in Eq. (5). It reveals the model's factualness.

$$Precision = \frac{1}{N}\sum\nolimits_{i=1}^{N} \frac{(TP)_i}{(TP + FP)_i} \tag{5}$$

When the dataset is unbalanced, F1-score is frequently used to show how accurate each class is. This study uses the F1-score measure to validate the fullness of the automatic speaker identification models because the datasets used in this work are unbalanced, making it difficult to calculate the accuracy of all speaker classes. Eq. (6) defines the F1-score as the weighted harmonic mean of recall and precision.

$$F1 - Score = \frac{1}{N}\sum\nolimits_{i=1}^{N} 2 \times \frac{(recall \times precision)_i}{(recall + precision)_i} \tag{6}$$

4 Results and Discussion

For automatic speaker identification, this study evaluated the proposed model by employing three datasets for text-independent experiments. The percentage spilt method was used to evaluate the proposed models, where 80% data was utilized to train the transformer model and the remaining 20% of data was utilized to test the trained model [30,31]. Previous works have concluded that the optimum performance of the model is obtained if the 20%–30% of the feature set is utilized for testing and the rest of 70%–80% feature set is used for training. The model achieves an accurate and valid accuracy for this data split method and does not report the overestimated accuracy.

The performance of the proposed transformer approach was evaluated using the ELSDSR, CSTR-VCTK, and Ar-DAD datasets. The proposed model obtained 100% training accuracy for each dataset except CSTR-VCTK where the proposed model obtained 98.5% training accuracy. This study's proposed transformer model increased precision and decreased loss for samples of training and testing data, demonstrating the utility and importance of the transformer method across all three datasets, as demonstrated in Fig. 5.

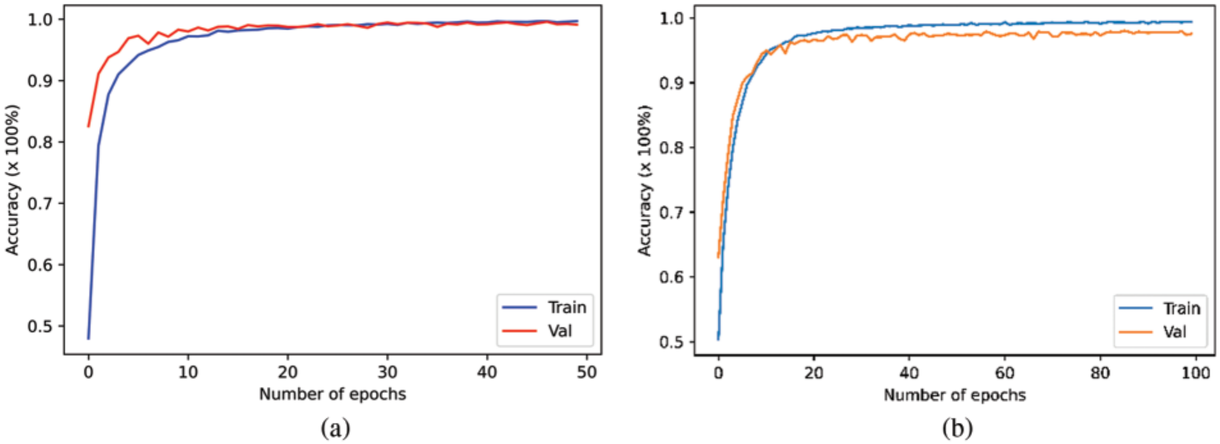
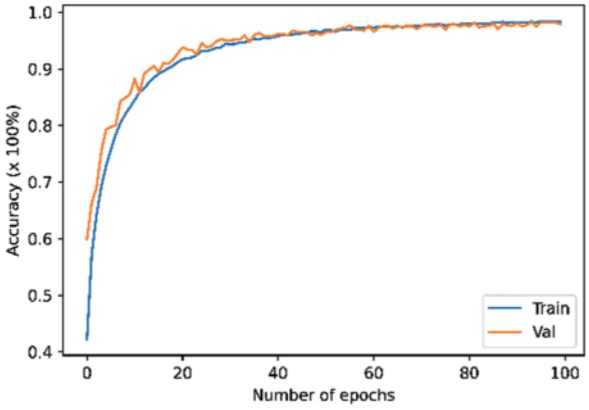


Figure 5: (Continued)



(c)

Figure 5: The proposed model training and validation accuracy for (a) Ar-DAD (b), CSTR-VCTK, and (c) ELSDSR

4.1 Models Prediction Performance

This research employed three datasets that have diverse speech signals. The prediction performance of all the conducted experiments are presented in Tables 2–4. These tables present the performance of proposed deep learning models for ELSDSR, Ar-DAD, and CSTR-VCTK. The tables reveal the robustness of the DL models over the baseline methods.

Table 2: Performance of proposed DL models for ELSDSR dataset

Deep learning model	Achieved performance			
	Accuracy	Precision	Recall	F-measure
Conventional CNN	0.97	0.97	0.97	0.97
TL (VGG−16)	0.97	0.96	0.97	0.97
Transformer	0.99	0.99	0.99	0.99

Table 3: Performance of proposed DL models for Ar-DAD dataset

Deep learning model	Achieved performance			
	Accuracy	Precision	Recall	F-measure
Conventional CNN	0.97	0.97	0.97	0.97
TL (VGG−16)	0.93	0.94	0.93	0.94
Transformer	0.99	0.99	0.99	0.99

Table 4: Performance of proposed DL models for CSTR-VCTK dataset

Deep learning model	Achieved performance			
	Accuracy	Precision	Recall	F-measure
Conventional CNN	0.92	0.92	0.92	0.92
TL (VGG−16)	0.95	0.95	0.95	0.95
Transformer	0.97	0.98	0.97	0.98

4.2 Comparison with Baseline Methods

This work used ELSDSR, Ar-DAD, and CSTR-VCTK to compare the results of the proposed model with the performance of existing baseline approaches to show the value and robustness of the suggested method for automatic speaker identification. Table 5 provides a thorough summary of the comparative analysis. The robustness of the suggested models is shown in the table, which shows that the proposed speaker identification models perform considerably better than baseline methods. In a few instances, the identification rate of the suggested models for specific speakers is slightly lower than the standard operating procedures. Although this effort obtained an average accuracy of 98% as opposed to the baseline of 93%, baseline procedures were exceeded by this work. The suggested ASI technique accurately identified each speaker, required less time and computer resources, and worked well for real-time applications.

Table 5: Comparison of the proposed method and baselines methods

Study	Dataset	Features	Model	Accuracy
[22]	ELSDSR	MFCCT	DNN	89%
[32]	ELSDSR	MFCC	ANN	93%
[33]	ELSDSR	MFCC	VQ-UBM-GMM	98%
Proposed model	ELSDSR	Chromagram, MFCCs, ΔMFCCs, ΔΔMFCCs, Spectral contrast, Mel-spectrogram, Tonnetz	Transformer	99%

4.3 Comparative Analysis with Machine Learning Classifiers

To build the automatic speaker recognition models, the resulting features were provided as input to five machine learning algorithms: Support Vector Machine (SVM), decision tree (J48), random forest (RF), Naive Bayes (NB), and k-Nearest Neighbour (k-NN). Additionally, 15 analyses (3 datasets × 5 ML techniques) were conducted to gauge how well the retrieved features and ML algorithms worked together.

Fig. 6 illustrates how the suggested transformer model beat all five machine learning techniques, with average accuracy for ELSDSR, Ar-DAD, and CSTR-VCTK of 98.8%, 99.6%, and 97.4%, respectively. The ML algorithms' average accuracy revealed a strange tendency. Using the ELSDSR dataset, the k-NN and RF algorithms achieved high weighted accuracy of 90.2% and 86.2%, respectively, as opposed to 79.2% for SVM, 76.4% for J48%, and 63.1% for NB. Furthermore, employing Ar-DAD and CSTR-VCTK, RF beat the other four ML methods. The J48 and SVM algorithms and the NB algorithm have the lowest levels of accuracy. In conclusion, employing all three datasets, the suggested transformer model for SER beat the ML classifiers because of its higher average accuracy.

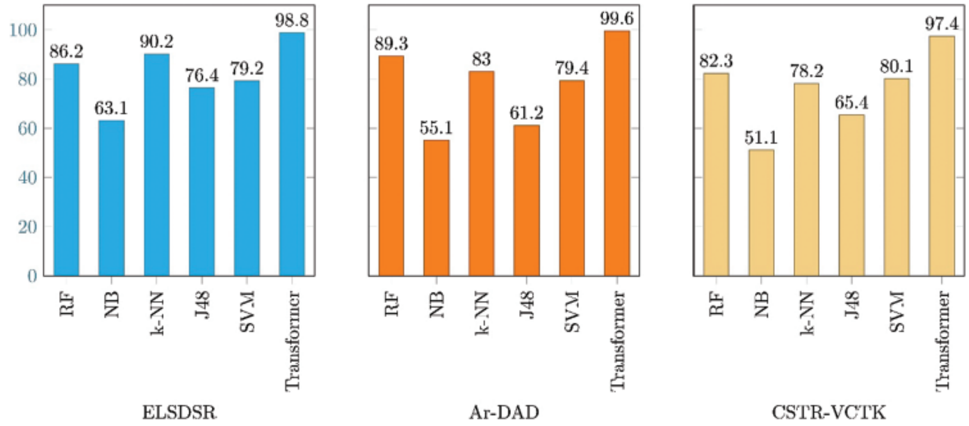


Figure 6: Comparative analysis of ML algorithms with proposed transformer model

5 Conclusion

Extraction of salient features and classification are two complex tasks in automatic speaker identification. A lightweight transformer model for speaker identification was presented in this study based on the combination of seven different acoustic properties. The effectiveness of the proposed technique was evaluated using the ELSDSR, Ar-DAD, and CSTR-VCTK datasets to demonstrate its robustness and importance. The proposed technique outperformed the baseline methods regarding recognition rate across all three datasets. Our proposed model achieved 99%, 97% and 99% accuracies for the ELSDSR, CSTR-VCTK and Ar-DAD datasets.

Speech analysis has recently drawn more and more attention. Numerous research challenges—from dataset collection methods to the usage of features—have emerged with the study of spontaneous human behavior (e.g., lexical information aside from prosodic features). However, there is little discussion of the function of contextual data in ASI. Exploring the role of contextual information can therefore be a helpful study contribution because it is clear that speaker identification depends heavily on the context.

Funding Statement: The authors are grateful to the Taif University Researchers Supporting Project Number (TURSP-2020/36), Taif University, Taif, Saudi Arabia.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] W. Lee, J. J. Seong, B. Ozlu, B. S. Shim, A. Marakhimov *et al.,* "Biosignal sensors and deep learning-based speech recognition: A review," *Sensors*, vol. 21, no. 4, pp. 1399, 2021.

[2] K. Simonyan, H. Ackermann, E. F. Chang and J. D. Greenlee, "New developments in understanding the complexity of human speech production," *Journal of Neuroscience*, vol. 36, no. 45, pp. 11440–11448, 2016.

[3] S. Lacey, Y. Jamal, S. M. List, K. McCormick, K. Sathian *et al.,* "Stimulus parameters underlying sound-symbolic mapping of auditory pseudowords to visual shapes," *Cognitive Science*, vol. 44, no. 9, pp. e12883, 2020.

[4] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (haspi) version 2," *Speech Communication*, vol. 131, pp. 35–46, 2021.

[5] W. T. Fitch, "The biology and evolution of speech: A comparative analysis," *Annual Review of Linguistics*, vol. 4, pp. 255–279, 2018.

[6] J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J. -F. Bonastre *et al.,* "Forensic speaker recognition," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 95–103, 2009.

[7] G. S. Morrison, F. H. Sahito, G. Jardine, D. Djokic, S. Clavet *et al.,* "INTERPOL survey of the use of speaker identification by law enforcement agencies," *Forensic Science International*, vol. 263, pp. 92–100, 2016.

[8] A. K. Hunt and T. B. Schalk, "Simultaneous voice recognition and verification to allow access to telephone network services," *Acoustical Society of America Journal*, vol. 100, no. 6, pp. 3488, 1996.

[9] J. Naik and G. Doddington, "Evaluation of a high performance speaker verification system for access control," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Dallas, TX, USA, pp. 2392–2395, 1987.

[10] M. Faundez-Zanuy, M. Hagmüller and G. Kubin, "Speaker identification security improvement by means of speech watermarking," *Pattern Recognition*, vol. 40, no. 11, pp. 3027–3034, 2007.

[11] C. Schmandt and B. Arons, "A conversational telephone messaging system," *IEEE Transactions on Consumer Electronics*, vol. 30, no. 3, pp. 21–24, 1984.

[12] J. C. Liu, F. Y. Leu, G. L. Lin and H. Susanto, "An MFCC-based text-independent speaker identification system for access control," *Concurrency and Computation: Practice and Experience*, vol. 30, no. 2, pp. 4255, 2018.

[13] F. Ye and J. Yang, "A deep neural network model for speaker identification," *Applied Sciences*, vol. 11, no. 8, pp. 3603, 2021.

[14] R. Jahangir, Y. W. Teh, F. Hanif and G. Mujtaba, "Deep learning approaches for speech emotion recognition: State of the art and research challenges," *Multimedia Tools and Applications*, vol. 80, no. 16, pp. 23745–23812, 2021.

[15] Z. Bai and X. -L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, 2021.

[16] A. Shafik, A. Sedik, B. Abd El-Rahiem, E. -S. M. El-Rabaie, G. M. El Banby *et al.,* "Speaker identification based on radon transform and CNNs in the presence of different types of interference for robotic applications," *Applied Acoustics*, vol. 177, pp. 107665, 2021.

[17] N. N. An, N. Q. Thanh and Y. Liu, "Deep CNNs with self-attention for speaker identification," *IEEE Access*, vol. 7, pp. 85327–85337, 2019.

[18] M. K. Nammous, K. Saeed and P. Kobojek, "Using a small amount of text-independent speech data for a BiLSTM large-scale speaker identification approach," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 3, pp. 764–770, 2022.

[19] A. B. Nassif, N. Alnazzawi, I. Shahin, S. A. Salloum, N. Hindawi *et al.,* "A novel RBFNN-CNN model for speaker identification in stressful talking environments," *Applied Sciences*, vol. 12, no. 10, pp. 4841, 2022.

[20] A. B. Nassif, I. Shahin, S. Hamsa, N. Nemmour and K. Hirose, "CASA-based speaker identification using cascaded GMM-CNN classifier in noisy and emotional talking conditions," *Applied Soft Computing*, vol. 103, pp. 107141, 2021.

[21] N. Maghsoodi, H. Sameti, H. Zeinali and T. Stafylakis, “Speaker recognition with random digit strings using uncertainty normalized HMM-based i-vectors,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1815–1825, 2019.

[22] R. Jahangir, Y. W. Teh, N. A. Memon, G. Mujtaba, M. Zareei *et al.,* “Text-independent speaker identification through feature fusion and deep neural network,” *IEEE Access*, vol. 8, pp. 32187–32202, 2020.

[23] S. Nainan and V. Kulkarni, “Enhancement in speaker recognition for optimized speech features using GMM, SVM and 1-D CNN,” *International Journal of Speech Technology*, vol. 24, no. 4, pp. 809–822, 2021.

[24] L. Sun, T. Gu, K. Xie and J. Chen, “Text-independent speaker identification based on deep Gaussian correlation supervector,” *International Journal of Speech Technology*, vol. 22, no. 2, pp. 449–457, 2019.

[25] H. Ali, S. N. Tran, E. Benetos and A. S. d’Avila Garcez, “Speaker recognition with hybrid features from a deep belief network,” *Neural Computing and Applications*, vol. 29, no. 6, pp. 13–19, 2018.

[26] P. Domingos, “A few useful things to know about machine learning,” *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[27] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou *et al.,* “Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network,” in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, pp. 5200–5204, 2016.

[28] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar *et al.,* “Librosa: Audio and music signal analysis in python,” in *Proc. of the 14th Python in Science Conf.*, Austin, TX, USA, pp. 18–25, 2015.

[29] R. Jahangir, Y. W. Teh, H. F. Nweke, G. Mujtaba, M. A. Al-Garadi *et al.,* “Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges,” *Expert Systems with Applications*, vol. 171, pp. 114591, 2021.

[30] E. Garcia-Ceja, M. Riegler, A. K. Kvernberg and J. Torresen, “User-adaptive models for activity and emotion recognition using deep transfer learning and data augmentation,” *User Modeling and User-Adapted Interaction*, vol. 30, no. 3, pp. 365–393, 2019.

[31] W. Nie, M. Ren, J. Nie and S. Zhao, “C-GCN: Correlation based graph convolutional network for audio-video emotion recognition,” *IEEE Transactions on Multimedia*, vol. 23, no. 1, pp. 3793–3804, 2020.

[32] M. Soleymanpour and H. Marvi, “Text-independent speaker identification based on selection of the most similar feature vectors,” *International Journal of Speech Technology*, vol. 20, no. 1, pp. 99–108, 2017.

[33] B. Barai, T. Chakraborty, N. Das, S. Basu and M. Nasipuri, “Closed-set speaker identification using VQ and GMM based models,” *International Journal of Speech Technology*, vol. 25, no. 1, pp. 173–196, 2022.