SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V13I9P113

Volume 13 Issue 9, 208-223, September 2026

Original Article

ECAPA-TF: A Hybrid ECAPA-Transformer Framework for Multilingual Speaker Identification using the NISP Corpus

Muliya Pritibahen Jagjivanbhai[1] , Ashwin Patni[2]

[1]*Indus University, Ahmedabad, Gujarat, India*
[2]*Department of Computer Science Engineering, Indus University, Ahmedabad, Gujarat, India*

[1]*Corresponding Author : pritibahenmuliya.23.rs@indusuni.ac.in*

Received: 30 May 2026 Revised: 07 August 2026 Accepted: 19 August 2026 Published: 29 September 2026

Abstract *- Speaker identification in multiple languages has received increasing attention in speech processing because of the growing use of voice-based security applications in multilingual settings. Traditional speaker identification systems have encountered difficulties ensuring reliable performance in bilingual and cross-language speech scenarios due to their concentration on local acoustic models and neglecting long-range contextual information. This research presents a novel hybrid architecture referred to as ECAPA-TF, which combines ECAPA-TDNN-based local acoustic modelling with transformer-based global contextual modelling for multilingual speaker identification. The proposed model is trained using the Native and Indian Speakers of Indian Phonetics (NISP) database consisting of 345 bilingual speakers of India, who are proficient in both English and native Indian languages, including Malayalam, Hindi, Kannada, Telugu and Tamil. The first step involves extracting Mel-Frequency Cepstral Coefficient (MFCC) feature vectors (40-dimensional) from the speech signal and normalizing the data prior to applying ECAPA residual blocks equipped with squeeze-excitation attention. Subsequently, the Transformer self-attention mechanism is implemented to identify long-range temporal and contextual dependencies in speech data. Attention pooling and Additive Angular Margin SoftMax (AAM-SoftMax) are also leveraged to create discriminative speaker embeddings. Performance is evaluated using Accuracy, recall, precision, confusion matrix, F1-score and ROC-AUC. The developed ECAPA-TF model achieved an accuracy of 92.93%, precision of 93.64%, recall of 93.02%, macro F1-score of 92.94%, and an AUC of 0.965. Under the same NISP experimental protocol, the proposed model outperformed the internally implemented Deep 1D-CNN, CNN-LSTM, and ECAPA baseline models. Compared with the ECAPA baseline, ECAPA-TF improved accuracy by 1.24 percentage points and macro F1-score by 1.36 percentage points. The confusion matrix and per-speaker evaluation further validate the effectiveness and reliability of the proposed approach under various multilingual speech scenarios.*

Keywords *- Multilingual speaker identification, Deep learning, ECAPA-TF, MFCC, AAM-SoftMax, NISP Corpus.*

1. Introduction

Speaker identification determines the identity of an unknown speaker by comparing their voice print with a list of enrolled and well-known users. Also, the distinction exists between what people call authentication (or speaker verification) and the broader identification part. In speaker recognition, both identifying and verifying who a human is, the approach really relies on multiple characteristics that the human voice signal can reveal, like acoustical information (short-time acoustic features) and prosody, such as the pitch period, stress, speech pace and other cues [1]. While developing a speaker recognition system, the speech signal's acoustic features are commonly used. Those features are expected to be effective, information with low redundancy, and still minimal in relation to the original size of the speech waveform [2].

Automatic verification or identification due to speaker recognition is possible for individuals purely by voice characteristics, and it serves as an important biometric technique for security and authentication systems [3]. It has been used in several applications, like mobile device authentication, forensic voice matching, telephone banking and also access control [4, 5].

With the fast development of Artificial Intelligence Technology, Automatic Speech Recognition (ASR) technology has now become an integral component of our day-to-day lives. ASR technology has a wide variety of uses, from voice searching, voice commands, subtitling videos, to personal assistants for smart devices, making it widely used. The conventional form of ASR technology has been centred around the recognition of only one language at a time and is

not very efficient with respect to multiple speakers, which limits its practicality and accuracy [6]. Classical approaches to speaker recognition started with acoustic-phonetic-based techniques [7, 8], and proceeded to MFCC [9], perceptual linear prediction-based representations [10], to the influential GMM-UBM framework by [11, 12]. The introduction of i-vectors marked an important development [13], which allows projecting an utterance into a fixed dimensionality total variability subspace in combination with joint factor analysis [14]. Deep learning methods allowed the development of d-vectors [15] and x-vectors [16-18] using TDNNs, which were later enhanced with attention mechanisms like ECAPA-TDNN [19]. Training of those models is largely made possible due to large-scale corpora such as VoxCeleb [20, 21] and LibriSpeech [22]. Recent progress has also been made with self-supervised methods like Wav2Vec2 [23], pioneering contrastive learning in the domain, HuBERT [24], introducing masked prediction with clustered labels.

Recent deep-learning methods have improved speaker recognition through different feature representations and neural network architectures [23, 25-29]. Most deep-learning-based speaker recognition systems extract acoustic features from speech signals and subsequently use these features for speaker modelling and classification. Common representations include FFT-based features, spectrograms, and MFCCs. However, the performance of such systems can be affected by feature-extraction parameters, language variations, accent differences, and recording conditions. In addition, several deep-learning architectures require a large number of trainable parameters for modelling the nonlinear characteristics of speech [29].

Multilingual speaker identification introduces an additional challenge because the acoustic and phonetic characteristics of the same speaker may vary when different languages are spoken. Existing speaker-recognition architectures primarily focus on extracting local acoustic characteristics, while long-range temporal and contextual relationships in speech may not be sufficiently represented. This problem is particularly relevant for bilingual Indian speakers because comparatively limited research has been carried out using Indian multilingual speech databases [30]. Therefore, there is a need for an architecture that can jointly model local speaker-specific acoustic information and long-range contextual information under multilingual speech conditions. The proposed ECAPA-TF framework addresses this research gap by integrating ECAPA-based local acoustic modelling with Transformer-based global contextual modelling for multilingual speaker identification.

The proposed contributions of the study:

- Proposed a hybrid ECAPA-Transformer (ECAPA-TF) framework for multilingual speaker identification.
- Integrated ECAPA-based local acoustic modelling with Transformer-based global contextual learning.
- Captured multi-scale temporal speech features using squeeze-excitation enhanced ECAPA blocks.
- Applied Attention Pooling and AAM-SoftMax to generate discriminative speaker embeddings.
- Evaluated the framework using accuracy, macro F1-score, confusion matrix, and per-speaker analysis.

1.1. Organization of the Paper

The remainder of the paper is organized as follows. Section 2 reviews related work and presents the technical background. Section 3 describes the dataset, preprocessing, and feature-extraction procedure. Section 4 presents the proposed ECAPA-TF architecture. Section 5 describes the training protocol and evaluation criteria. Section 6 reports and discusses the experimental results. Section 7 concludes the study.

2. Related Work and Background

This section presents the existing models worked on multilingual speaker identification and their limitations. Moreover, provided the background of the proposed models.

2.1. Related Work

Experiments on the YOHO corpus produced a maximum average accuracy of 91.93%.

Almarshady et al. [2] developed a DNN-based speaker-identification system using acoustic and prosodic features, including pitch and energy. The study explores adaptations of existing RNN models for this system, employing the public YOHO LDC dataset, achieving an average accuracy of 91.93%. Due to MLP limitations, alternative approaches are considered, acknowledging the quasi-stationary and time-varying nature of speech signals. In [31], Singh et al. developed an Automatic Speaker Identification (ASI) system to classify native Hindi, Punjabi, and Bengali speakers by regional accents. Mel-Frequency Cepstral Coefficients (MFCC) and linear predictive cepstral coefficients are used to extract features from native speaker voice data. SVM and DT classifiers are used to analyze non-native English data impacted by these accents.

The SVM classifier outperforms the DT classifier with 96.23% vs 93.75%, showing that cepstral characteristics and SVM can detect Indian speakers' native accents. Nugroho et al. [32] presented a data augmentation method using adding white noise, time stretching, and Pitch Shifting to solve speaker recognition with a 7-layer deep neural network referred to as DA-DNN7L. The presented method addresses the data problem facing Indonesian ethnic speakers. Using a split of 70%:30% data containing 400 samples, a seven-layer deep neural network achieved 99.76% accuracy at a loss of 0.05. Vajrobol et al. [33] presented an Encoder-MFCC-Chroma STFT Hybrid model, which helps address the problem of multilingual speaker identification. The model is

efficient with an accuracy of 99.04% while dealing with a well-prepared dataset for multilingual speakers. Khmer achieves an accuracy of 98.16%, precision of 98.19%, and a recall of 97.91%, resulting in its F1 score of 97.95%. Chroma STFT and MFCC help achieve good speaker discrimination in low-resourced languages like Khmer, Javanese, Sudanese and Nepali. Although the method performs well in low-resource languages, the technique is constrained by the type of data set utilized in the training process.

Sritharan and Thayasivam [34] presented a novel multilingual speaker recognition and verification method for Indo-Aryan and Dravidian languages. In relation to the accuracy of recognizing speakers in comparison with monolingual models, the Kathbath dataset demonstrates an increase in the range of 15% to 1%. In the case of speaker verification, equal error rate in conditions of noise, there is a decrease in the range from 15% to 5%. Due to the diversity of the training dataset, i.e., Kathbath, the applicability of the dataset is limited, and model computational complexity may restrict its application in real-time or resource-limited settings.

In [35], Tao et al. find that a speaker recognition network learns from reliable labels more quickly than from unreliable ones. Based on this observation, Tao et al. [35] introduced a Loss-Gated Learning (LGL) strategy that prioritizes reliable labels during training. Implementing LGL results in a 46.3% performance improvement compared to models without it.

Additionally, the self-supervised speaker recognition model, trained with LGL on the VoxCeleb2 dataset, achieves an equal error rate of 1.66% on the VoxCeleb1 test set. For better information capture and performance, Li et al. [36] introduced the Efficient Parallel Channel Network - Time Delay Neural Network (EPCNet-TDNN), which replaces the SE_block with a unique Efficient Channel and Spatial Attention Mechanism (ECAM).

The model uses ECAPA-TDNN's Tandem Structure (TS) to integrate multi-scale and channel information and a Parallel Residual Structure (PRS) for parallel computing of multi-scale features. Additional temporal feature capture is provided by the Selective State Space (SSS) module. According to CN-Celeb1 dataset experiments, EPCNet-TDNN improves EER by 14.1%, minDCF by 9.4%, and ACC by 6.6% over ECAPA-TDNN.

Hanifa et al. [37] presented a methodology of speech recognition for identifying the ethnicity of people in Malaysia. From 52 continuous speeches of the Malay language, the pitch and 13 Mel-Frequency Cepstrum Coefficients (MFCCs) are selected as features to build the classification models based on the Tree, Naive Bayes, Nearest Neighbors, and SVM. Another 10 recordings are selected for testing. It is evident that using the combination of pitch and 13 MFCCs as features and building models based on SVM gives better classification performance (57.7%) compared to using 13 coefficients only (53.8%).

In [38], M. Ahmed et al. presented Neuro-TM Diarizer, a deep learning architecture for boosting performance in speaker diarization using noise cancellation, adaptive beamforming, and neural mechanisms. By using Marble-Net for speech detection and Tita-Net for speaker embedding, this architecture employs time-delay neural networks for recognition. The experiments on the VoxConverse and VoxCeleb datasets reveal that Neuro-TM Diarizer outperforms clustering algorithms substantially, with DERs of 6.89% and 6.93% and DER improvements of 12.60% and 14.01%, respectively.

Efat et al. [39] investigated feature extraction through signal processing and deep neural networks through advanced machine learning classifiers. It makes use of a combined GRU-CNN to maximise the subspace loss function for feature vector extraction. The VoxCeleb database, which contains data from 6,000 speakers, provided the highest accuracy and F1 score when the GRU-CNN and R-CNN combination was used, while the highest precision and recall were achieved by the CNN and LPC-KNN combinations, respectively.

In [40], Pan et al. proposed a new speaker identification model, Chrono-ECAPA-TDNN (CET), which improves accuracy in complicated noise and brief speech characteristics of civil aviation air-ground communication. Chrono Block extracts features using parallel branching and attention mechanisms; the speaker embedding extraction module uses self-attention pooling; and an optimised loss function with Sub-center AAM-Softmax improves class separation. Robustness is improved via data augmentation. Prétrained on VoxCeleb2 and evaluated on air-ground communication data, the model outperformed the baseline ECAPA-TDNN model by 1.53% and 2.19% in EER and accuracy, respectively.

Table 1. Summary of the existing studies

Ref.	Method	Dataset / Setting	Multilingual	Indian Data	Main Modelling Strength	Limitation Relative to Proposed Work
[2]	DNN	YOHO	No	No	Acoustic/prosodic modelling	Limited long-range contextual modelling
[31]	SVM/DT	Indian regional accent speech	No	Yes	Accent-based classification	Conventional ML; no deep contextual modelling

[32]	DA-DNN7L	Indonesian ethnic speech	No	No	Data augmentation + DNN	No multilingual contextual modelling
[33]	MFCC + Chroma STFT + Transforme r	Low-resource multilingual speech	Yes	No	Hybrid acoustic + Transformer representation	Not ECAPA-based
[34]	IndicWav2 Vec	Kathbath	Yes	Yes	Self-supervised multilingual representation	Different architecture and training paradigm
[35]	LGL	VoxCeleb	No	No	Reliable-label learning	Does not combine ECAPA and Transformer
[36]	EPCNet-TDNN	CN-Celeb1	No	No	Improved channel and temporal modelling	No Transformer-based global contextual block
[37]	ML classifiers	Malay ethnic speech	No	No	Pitch + MFCC classification	Conventional ML
[38]	Neuro-TM Diarizer	VoxConverse/VoxCe leb	Yes/Multispeak er	No	Diarization architecture	Targets diarization rather than identification
[39]	GRU-CNN	VoxCeleb	Limited	No	Hybrid temporal/convolution al modelling	No ECAPA-Transformer integration
[40]	Chrono-ECAPA-TDNN	Air-ground speech	No	No	ECAPA + self-attention pooling	Domain-specific; no Transformer global block
Propo sed	ECAPA-TF	NISP	Yes	Yes	ECAPA local + Transformer global modelling	-

The reviewed studies demonstrate substantial progress in speaker identification through deep neural networks, hybrid feature representations, self-supervised learning, TDNN-based architectures, attention mechanisms, and multilingual modelling [2, 31-40]. Conventional DNN and machine-learning approaches [2, 31, 32, 38] mainly rely on acoustic or handcrafted feature representations and provide limited modelling of long-range temporal dependencies.

Hybrid approaches such as MFCC-Chroma STFT [33] and GRU-CNN [39] improve feature diversity but do not specifically combine ECAPA-based multi-scale acoustic modelling with Transformer-based contextual learning. IndicWav2Vec [34] addresses multilingual Indian speech using a self-supervised representation, whereas LGL [35] improves training reliability through label-dependent loss control.

EPCNet-TDNN [36] and Chrono-ECAPA-TDNN [40] extend TDNN- and ECAPA-based modelling through improved attention and residual processing, but these approaches were evaluated on datasets and acoustic conditions different from the bilingual NISP setting considered in this study. Neuro-TM Diarizer [38] incorporates multiple deep-learning modules for diarization rather than closed-set multilingual speaker identification. These observations indicate that the existing literature addresses individual aspects of speaker modelling, but the joint use of ECAPA-based local acoustic representation and Transformer-based global temporal modelling for bilingual Indian speaker identification remains insufficiently explored.

2.2. Research Gap

The existing literature shows that speaker-identification performance has been improved through deep neural networks, self-supervised representations, hybrid feature extraction, TDNN-based models, and attention mechanisms [2, 31-40]. However, several limitations remain for multilingual speaker identification. First, many reported methods are evaluated on monolingual, general multilingual, ethnic, or domain-specific datasets rather than on bilingual Indian speech.

Second, TDNN- and ECAPA-based methods primarily emphasize local and multi-scale acoustic modelling, whereas Transformer-based approaches are better suited to modelling long-range temporal relationships. Third, only limited work has examined the joint use of these complementary modelling

mechanisms within a single architecture for closed-set identification of bilingual Indian speakers. The NISP corpus provides speech from speakers using English and one native Indian language, making it suitable for evaluating speaker representation under multilingual conditions.

Therefore, a model is required that can preserve local speaker-dependent acoustic characteristics while also modelling longer-range temporal relationships across speech frames. The proposed ECAPA-TF framework addresses this requirement by integrating ECAPA-style residual acoustic modelling with Transformer self-attention, followed by attention pooling and AAM-SoftMax-based speaker classification.

2.3. Novelty of the Proposed Work

The novelty of ECAPA-TF lies in the sequential integration of ECAPA-based multi-scale acoustic modelling and Transformer-based global contextual modelling for multilingual speaker identification on the NISP corpus. Unlike the reviewed DNN, machine-learning, hybrid feature, self-supervised, and ECAPA-based approaches [2, 31-40], the proposed architecture combines squeeze-excitation-enhanced ECAPA residual blocks, Transformer self-attention, learnable attention pooling, and AAM-SoftMax within a unified speaker-identification framework. This combination is intended to preserve short-range speaker-specific acoustic cues while incorporating long-range temporal relationships before generating the final speaker embedding.

The studies discussed in this section were conducted using different datasets, speaker populations, language conditions, and evaluation protocols. Therefore, their reported performance values are used only to provide methodological background and are not considered directly comparable with the results obtained on the NISP corpus. For a fair experimental comparison, the proposed ECAPA-TF model is evaluated against internally implemented ECAPA, Deep 1D-CNN, and CNN-LSTM baselines trained and tested under the same NISP data split and evaluation protocol.

2.4. Background

2.4.1. ECAPA-TDNN Model

ECAPA-TDNN is an improved iteration of the x-vector model that has exhibited impressive performance in various speaker verification competitions, notably VoxSRC-2019 [41], VoxSRC-2020 [42], SdSVC-2021 [43], and VoxSRC-2021 [44]. ECAPA-TDNN utilizes a one-dimensional convolutional layer, in which the input feature is a two-dimensional representation formatted as FxT, with F denoting the frequency dimension and T indicating the time dimension.

The input undergoes initial processing using a 1D inflated convolutional layer, yielding a feature map of size CT, with C being the total number of channels. The feature map undergoes processing through three successive SE-Res2Blocks, in which each consists of two 1D inflated convolutional layers followed by a 1D inflated Res2Block, and a 1D Squeeze-and-Excitation (SE) block. A multilayer feature aggregation is used to acquire and aggregate hierarchical speaker information from various network layers. It is done by connecting the output feature maps of the three SE-Res2Blocks.

Subsequently, an attention-based statistical pooling layer is utilized to highlight speaker-specific characteristics across both channel and temporal dimensions using a self-attention mechanism. Finally, akin to ResNet-based speaker verification systems, a fully connected layer is employed to diminish the dimensionality of the pooled vectors, while AAM-softmax functions as the loss criterion for model training. The model is illustrated in Figure 1.

3. Methodology

Speaker identification - the task of determining which enrolled individual produced a given utterance- remains a non-trivial problem when speech is collected across multiple languages and regional accents. Most benchmark corpora are monolingual, which limits what can be learnt about how language-switching or code-mixing behavior affects the acoustic characteristics that distinguish one person from another.

The NISP corpus, released by the LEAP Lab at the Indian Institute of Science, Bangalore, was assembled precisely to fill that gap. It covers 345 bilingual speakers, 219 male and 126 female, each of whom contributed roughly four to five minutes of read speech in two languages: English and one of five South or North Indian mother tongues (Hindi, Kannada, Malayalam, Tamil, and Telugu). This dual-language design makes NISP an ideal testbed for studying cross-lingual speaker consistency, that is, whether the cues that make a person recognizable persist even when they change language.

Against this backdrop, a model architecture is proposed that deliberately blends two complementary paradigms in sequence modelling. The local, channel-rich feature abstraction characteristic of the ECAPA-TDNN family is combined with the global contextual reasoning of a Transformer self-attention block. The resulting network is trained with AAM-SoftMax, an objective that explicitly pushes speaker embeddings apart in angular space.

The system was then evaluated in a closed-set identification setting, reporting accuracy, macro F1-score, confusion matrices, and per-speaker analysis. The following sections describe each component in detail, beginning with the corpus and concluding with a discussion of training protocol and evaluation strategy. Figure 2 shows the flowchart of the proposed work.

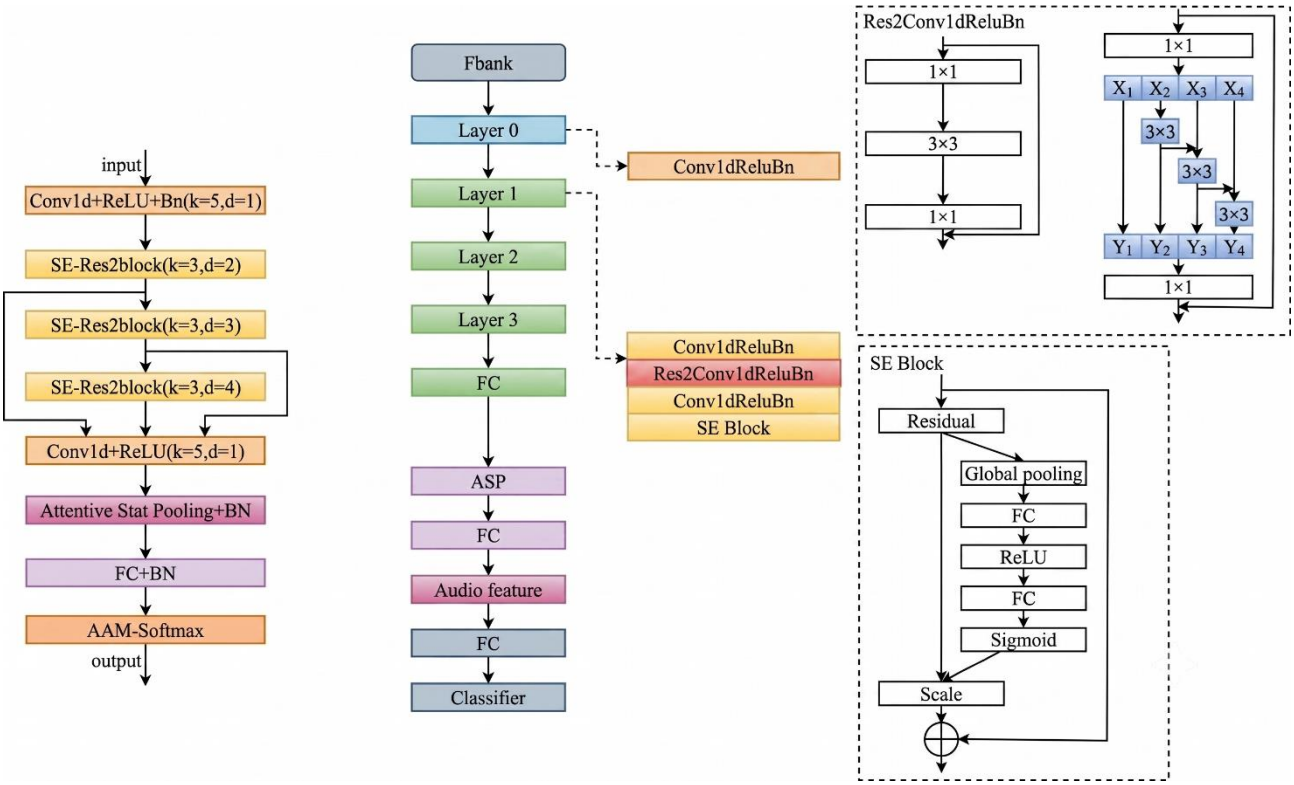

Fig. 1 Structure of ECAPA-TDNN [36]

3.1. Dataset: The NISP Corpus

3.1.1. Corpus Overview

The NISP corpus was introduced by Babu et al. [45] through IISc LEAP Lab and is publicly accessible at https://github.com/iiscleap/NISP-Dataset. Beyond speech recordings, NISP is notable for providing rich metadata for every speaker: biological sex, mother tongue, medium of instruction, age, height, weight, shoulder and waist measurements, and geographic origin down to district and state.

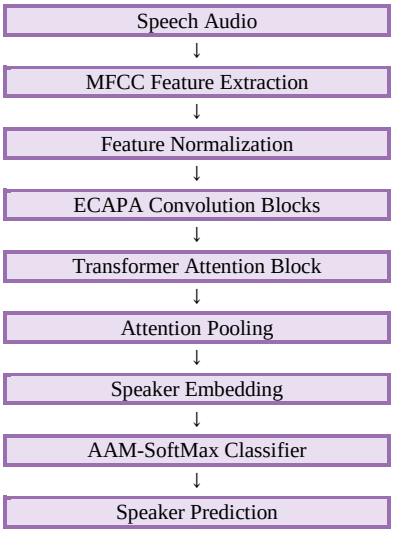

Fig. 2 Flowchart of the proposed work

The corpus is structured around five native language folders (Hindi, Kannada, Malayalam, Tamil, and Telugu), each containing two sub-directories: one for utterances spoken in the native tongue and one for English utterances recorded in a separate session. File names encode language prefix, a four-digit speaker ID, the recorded language, biological sex, and a zero-indexed utterance counter (e.g., Hin_0001_Eng_f_0000.wav). Transcripts are provided in UTF-8 text files aligned per speaker and language.

3.2. Bilingual Filtering and Speaker Splits

For the speaker identification task, the attention is restricted to the speakers who have recordings in at least two languages, referred to throughout this paper as bilingual speakers.

This ensures every enrolled identity has acoustic samples drawn from distinct linguistic contexts, which is the more challenging and more realistic scenario. The NISP repository supplies explicit train and test speaker ID lists (train_spkrID and test_spkrID).

In the experiments reported here, a closed-set protocol is adopted in which training and test utterances are drawn from the same pool of speakers but from non-overlapping audio segments partitioned within each speaker at a ratio of 80%/20% for training/testing, with a further 20% of the training portion held out as a validation set.

Table 2. Summary statistics of the NISP corpus used in this study

Property	Value
Total speakers	345
Male / Female	219 / 126
Languages	English + Hindi / Kannada / Malayalam / Tamil / Telugu
Approx. speech per speaker	4 - 5 minutes
Train / Test split	Closed-set, 80% / 20% within speaker
Audio format	16 kHz, 16-bit WAV
Metadata provided	Age, height, weight, native state/district

3.3. Acoustic Feature Extraction

3.3.1. Mel-Frequency Cepstral Coefficients

Raw waveforms sampled at 16 kHz are converted to Mel-Frequency Cepstral Coefficients, which remain among the most widely used and well-understood short-time spectral representations in speaker recognition. For each utterance, C = 40 MFCC dimensions are extracted, producing a matrix x ∈ R^(T×C) where T is the number of frames.

Frame length is 25 ms with a hop of 10 ms, yielding a frame rate of 100 Hz. Utterances are zero-padded or truncated to a fixed length of T_{max} = 200 frames, corresponding to two seconds of speech, long enough to capture several phoneme-level patterns but short enough to keep memory requirements manageable during batch training.

The choice of 40 MFCC dimensions reflects a balance: fewer coefficients reduce speaker discriminability while more coefficients risk capturing channel artefacts or microphone noise rather than pure vocal-tract characteristics. The final feature tensor passed to the neural network is therefore of shape (200, 40).

3.3.2. Feature Standardization

Prior to model training, per-feature z-score standardization is applied across the training set: for each of the 40 MFCC dimensions, the empirical mean μ_j and standard deviation σ_j are computed over all training frames. The same statistics are then used without re-estimation to normalize validation and test frames. Formally, for frame vector x ∈ R^C, the normalized vector is

$$\hat{x}_j = \frac{x_j - \mu_j}{\sigma_j + \varepsilon} \quad (1)$$

Where ε = 10^{-7} prevents division by zero. This normalization step is critical in a multilingual setting because different language sessions may be recorded with slightly different microphone gains or room acoustics, and centring the features reduces such session-level bias without requiring explicit channel compensation.

4. Proposed Model Architecture

The proposed network, which is termed ECAPA-Transformer (ECAPA-TF), is designed as a three-stage pipeline: a local convolutional front-end inspired by ECAPA-TDNN, a global Transformer reasoning block, and a discriminative embedding layer trained with AAM-SoftMax. Figure 3 illustrates the complete flow.

4.1. Convolutional Front-End

The input MFCC sequence is first projected from 40 dimensions into a higher-dimensional feature space by a single Conv1D layer with 128 filters, kernel size 5, same- padding, no bias, followed by Batch Normalization and ReLU activation. This layer acts as a learned frequency-domain filter bank that aggregates local context across approximately 50 ms (five frames at 10 ms hop).

4.2. ECAPA Res2 Blocks with Squeeze-Excitation

Three successive ECAPA-style residual blocks operate on the projected features. Each block takes an input tensor H ∈ R^(T×192) and produces a transformed output through the following sequence. Let d ∈ {1, 2, 3} denote the dilation rate assigned to the k-th block.

Step 1 Bottleneck conv: A pointwise Conv1D (kernel = 1) projects the channels, followed by BN and ReLU.

Step 2 Dilated depth-wise conv: A Conv1D with kernel size 3 and dilation rate d expands the temporal receptive field without increasing parameter count. With dilation d, the effective receptive field covers 2d + 1 consecutive frames.

Step 3 Squeeze-Excitation gate: Channel attention is applied via a Squeeze-Excitation (SE) block:

$$s = sigmoid\left(W^2 \cdot ReLU\left(W^1 \cdot GAP(H')\right)\right) \quad (2)$$

Where GAP denotes Global Average Pooling along the time axis, $W_1 \in R^{(r \times C)}$ and $W_2 \in R^{(C \times r)}$ are the squeeze and excitation weights, respectively, with reduction ratio r = ⌊C/8⌋.

The scale vector s ∈ R^C is reshaped to (1, C) and broadcast-multiplied element-wise along the time dimension. This mechanism lets each block focus on the most informative frequency-like channels for each temporal region.

Step 4 Residual add: The SE-gated output is added to the identity shortcut (projected to 192 channels if dimensions differ) and passed through a final ReLU activation.

The use of three blocks with progressively larger dilation rates (1, 2, 3) allows the network to capture acoustic events at multiple temporal scales simultaneously, from phone-level detail in block 1 to syllable-level rhythm in block 3.

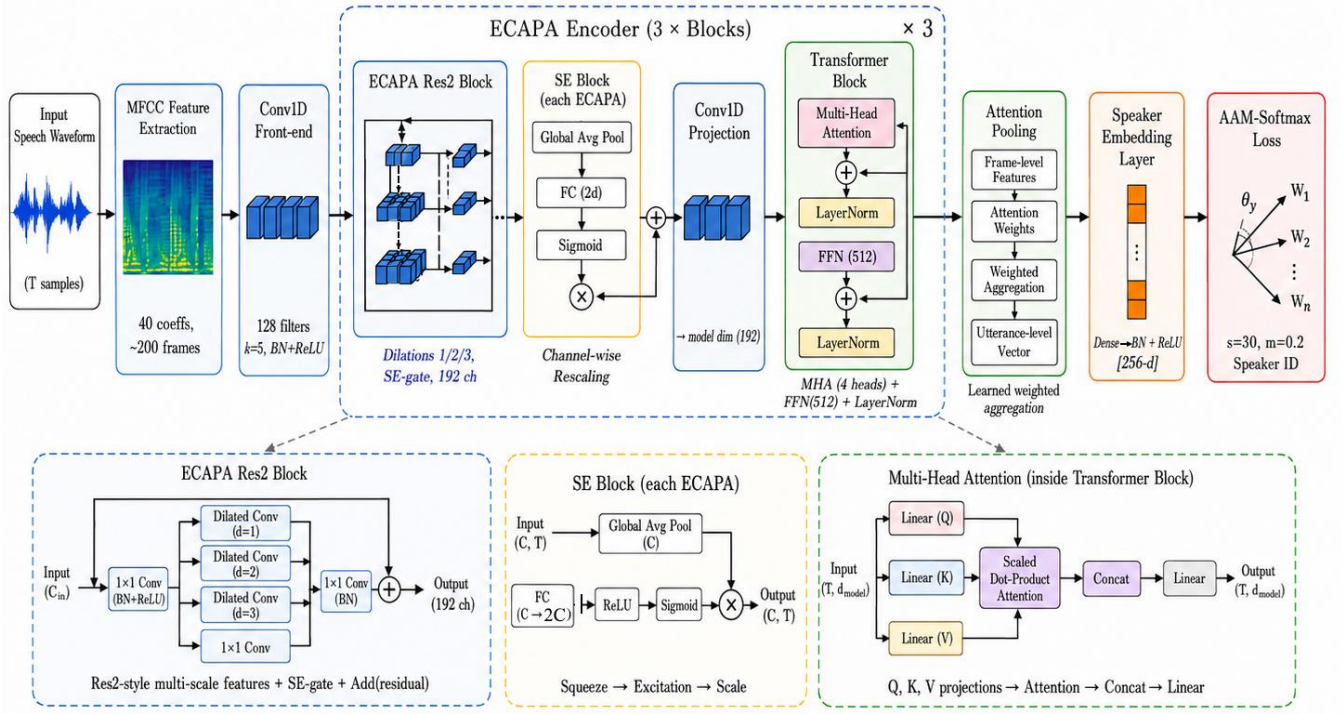

Fig. 3 Architecture of the proposed ECAPA-TF speaker identification network

4.3. Transformer Self-Attention Block

After the ECAPA front-end, the sequence is projected to D = 192 dimensions via a pointwise Conv1D and then fed to a single Transformer block, following the standard formulation of Vaswani et al. [46]. The block consists of a Multi-Head Attention (MHA) sub-layer and a position-wise Feed Forward Network (FFN), each wrapped with a residual connection and Layer Normalization. Multi-Head Attention computes scaled dot-product attention independently across H = 4 heads:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (3)$$

where Q, K, V are linear projections of the input, where $d_k = D/H = 48$ is the per-head dimension. Then outputs of all heads are concatenated and projected to D dimensions. The FFN applies two fully-connected layers with a ReLU nonlinearity:

$$FFN(x) = ReLU(xW^1 + b^1)W^2 + b^2 \qquad (4)$$

with inner dimension $d^{\text{ɔm}} = 512$.

A Dropout rate of p = 0.1 is applied within both the attention weights and the FFN. The Transformer block allows any frame to attend directly to any other frame in the utterance, capturing long-range co-articulation patterns and speaker-specific prosodic structure that local convolutions alone cannot reach.

4.4. Attention Pooling

To compress the variable-time-length sequence into a single fixed-dimensional vector, simple mean- or max-pooling is replaced with a learnable Attention Pooling layer. Given the Transformer output H ∈ R^(T×D), a small sub-network computes a scalar importance weight for each frame:

$$p^t = H^t W_{proj} \qquad (5)$$

$$\alpha^t = softmax(tanh\ tanh\ (p^t)\ w_{attn}) \qquad (6)$$

$$e = \Sigma_t\ \alpha^t \cdot p^t \qquad (7)$$

where W_{proj} ∈ R^(D×D) and W_{attn} ∈ R^(D×1) are learnable parameters. The output e ∈ R^D is an utterance-level summary that emphasizes frames deemed most informative by the attention mechanism, naturally de-emphasizing silent or noisy regions without requiring explicit voice activity detection.

4.5. Speaker Embedding Layer

The pooled vector e is applied through a fully-connected layer to produce a 256-D speaker embedding, followed by Batch Normalization and ReLU. Formally:

$$z = ReLU(\ BN(\ e\ W\_emb\ +\ b\_emb\ )\ ) \qquad (8)$$

This 256-d vector z constitutes the speaker embedding and can be used post-training for tasks such as speaker verification or clustering. It is analogous to the x-vector or d-

vector used in conventional speaker recognition pipelines, but produced end-to-end by the ECAPA-TF network.

4.6. Additive Angular Margin SoftMax (AAM-SoftMax)

Speaker recognition benefits from objectives that enforce large inter-class angular margins in the embedding space. Vanilla SoftMax cross-entropy does not explicitly penalize embeddings that are directionally close to multiple class centres. AAM-SoftMax, introduced by Deng et al. (2019) as Arc Face in the face recognition context and subsequently adopted for speaker recognition, addresses this by adding a fixed angular margin m to the angle between the embedding and its target class prototype.

5. Training Protocol

5.1. Optimizer and Schedule

The full ECAPA-TF network is trained end-to-end with the Adam optimizer using $\beta_1 = 0.9$, $\beta_2 = 0.999$, a learning rate of 1×10^{-4}, and weight decay disabled. Training proceeds for up to 50 epochs having a batch size of 64. Early stopping is applied on validation loss (with a patience of seven epochs), restoring the best weights observed during training. The relatively modest learning rate, combined with early stopping, guards against overfitting on the limited per-speaker data available in NISP without requiring a warm-up schedule.

5.2. Data Pipeline

Mini-batches are assembled from the TensorFlow tf.data API with shuffling and prefetching. Each sample is a tuple (mfcc_tensor, speaker_label) where the tensor has already been padded or cropped to 200 frames. No online data augmentation (speed perturbation, Spec Augment, additive noise) was applied in the baseline experiments reported here; this is a direction for future improvement.

5.3. Baseline models: ECAPA, Deep 1D-CNN, CNN-LSTM

To provide a fair comparison with the proposed ECAPA-TF model, three neural baselines, namely ECAPA, Deep 1D-CNN, and CNN-LSTM, are evaluated using the same closed-set NISP split and MFCC-based preprocessing.

The baseline networks are implemented in TensorFlow/Keras using standardized frame-level MFCC inputs. Training is performed with the Adam optimizer and sparse categorical cross-entropy loss, while validation loss is monitored using early stopping with restoration of the best model weights. During testing, class scores are generated for each speaker and the class with the maximum score is selected as the predicted identity.

5.4. Evaluation Metrics

Model performance is quantified using the following metrics, each computed on held-out test utterances that were never seen during training or validation.

5.5. Accuracy

Closed-set identification accuracy is defined as the fraction of test utterances for which the predicted speaker identity matches the ground truth: $\text{Acc} = (1/N)\ \Sigma\ I[\hat{y}_i = y_i]$, where N is the total number of test samples.

Macro F1-Score: Because speaker counts are not always perfectly balanced, accuracy alone can mask poor performance on less-represented speakers. Macro F1-score averages the per-class F1 without weighting by support, giving equal importance to each enrolled speaker regardless of how many utterances they contributed.

Confusion Matrix: A normalized confusion matrix is produced for visual inspection of systematic errors. Off-diagonal clusters where utterances from one speaker are consistently misidentified as another often point to shared regional accents or recording conditions, which is particularly relevant in a multilingual corpus such as NISP.

Per-Speaker Analysis: In addition to aggregate figures, per-speaker precision and recall are recorded and stored alongside individual audio paths, enabling post-hoc analysis of which speakers or language conditions prove most challenging for the system.

5.6. Implementation Details

Table 3. Hyperparameters and implementation settings for all experiments

Hyperparameter	Value
Sample rate	16 000 Hz
MFCC coefficients (C)	40
Sequence length (T_max)	200 frames
ECAPA channel width	192
Transformer model dim (D)	192
Transformer heads (H)	4
Transformer FFN width	512
Embedding dimension	256
AAM scale (s)	30.0
AAM margin (m)	0.2
Dropout (p)	0.1
Optimizer	Adam
Learning rate	1×10^{-4}
Batch size	64
Max epochs	50
Early stopping patience	7 epochs
Feature cache	Disk-backed NumPy arrays
Deep learning framework	TensorFlow / Kera's
Baseline models	ECAPA, Deep 1D-CNN, CNN-LSTM

Algorithm

Algorithm 1: ECAPA-TF Based Multilingual Speaker Identification

Input:

NISP multilingual speech dataset D
Speaker labels Y

Output:

Predicted speaker identity $\hat{Y}$
Speaker embeddings Z
Evaluation metrics

Begin

1: Load NISP speech dataset D
2: Select bilingual speaker utterances
3: Split the dataset into training, validation, and testing sets
4: **For** each audio sample ai ∈ D do
5: Resample audio to 16 kHz
6: Apply framing and windowing
7: Extract 40-dimensional MFCC features
8: Pad/Truncate sequence to $T_{max} = 200$ frames
9: **End For**
10: Compute feature mean μ and standard deviation σ
11: Normalize MFCC features using: $\hat{x}_j = \frac{x_j - \mu_j}{\sigma_j + \varepsilon}$
12: Initialize ECAPA-TF network
13: Pass normalized MFCC features into Conv1D layer
14: Apply Batch Normalization and ReLU activation
15: **For** each ECAPA residual block, do
16: Apply bottleneck convolution
17: Apply dilated convolution
18: Apply squeeze-excitation attention
19: Apply residual connection
20: **End For**
21: Project features into Transformer embedding space
22: Apply Multi-Head Self-Attention:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

23: Apply Feed Forward Network and Layer Normalization
24: Compute attention pooling weights
25: Generate utterance-level feature representation
26: Pass pooled features through a fully connected layer
27: Generate 256-dimensional speaker embedding Z
28: Apply AAM-SoftMax loss function
29: Predict speaker identity $\hat{Y}$
30: Train network using Adam optimizer
31: Validate model using validation dataset
32: Apply early stopping if validation loss stagnates
33: Evaluate model using test dataset
34: Compute:
Accuracy
Macro F1-score
Confusion Matrix
Per-speaker analysis
35: Compare results with baseline
End

The experiments were conducted using a fixed train-validation-test partition. The test set contained 5504 utterances from the 345 enrolled speakers. All reported performance measures correspond to a single experimental run using the fixed data partition and model initialization adopted for the study. Therefore, the reported accuracy, precision, recall, F1-score, and ROC-AUC values represent point estimates from that run, and mean ± standard deviation values across multiple seeds are not reported.

6. Results and Discussion

The proposed ECAPA-TF architecture was evaluated on the NISP corpus using a closed-set identification protocol involving 345 speakers. Performance was compared with ECAPA, Deep 1D-CNN, and CNN-LSTM baselines using accuracy, precision, recall, macro F1-score, confusion matrices, and ROC-AUC. To further understand the classification characteristics and generalizability of all four models, confusion matrices, ROC-AUC analysis and convergence graphs have been provided.

6.1. Overall Performance Metrics

To ensure a fair comparison, all baseline models reported in Table 4 were implemented and evaluated using the same NISP corpus, speaker partitioning strategy, preprocessing procedure, and evaluation metrics as the proposed ECAPA-TF model. The proposed ECAPA-TF model achieved an accuracy of 92.93%, precision of 93.64%, recall of 93.02%, and macro F1-score of 92.94%. Its accuracy exceeded that of Deep 1D-CNN and CNN-LSTM by 10.17 and 11.61 percentage points, respectively. Compared with the ECAPA baseline, ECAPA-TF improved accuracy by 1.24 percentage points and macro F1-score by 1.36 percentage points.

Table 4. Classification performance comparison of the proposed ECAPA-TF model with internally implemented baselines on the NISP dataset (345 speakers)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Proposed Model	92.93	93.64	93.02	92.94
ECAPA	91.69	92.65	91.76	91.58
Deep 1D-CNN	82.76	85.28	82.85	82.82
CNN-LSTM	81.32	83.47	81.57	81.50

A recall rate of 93.02% shows that the model recognizes the majority of the 345 speakers, and an accuracy level of 92.93% shows that the model makes few mistakes when identifying the speaker.

6.2. Speaker-Level Recall and Consistency

The speaker-level Macro Recall and standard deviation (σ) for each of the 345 speakers are displayed in Table 5. For various speaker characteristics, a lower standard deviation indicates more dependable results-a crucial quality required for real-world applications.

Table 5. Speaker-level macro recall and standard deviation across 345 speakers on the NISP dataset

Model	Macro Recall	Std. Deviation (σ)
Proposed Model	0.9302	0.0973
ECAPA	0.9176	0.1175
Deep 1D-CNN	0.8285	0.1483
CNN-LSTM	0.8157	0.1427

The proposed model produced the highest macro speaker-level recall of 0.9302 and the lowest standard deviation of 0.0973. The corresponding standard deviations for ECAPA, Deep 1D-CNN, and CNN-LSTM were 0.1175, 0.1483, and 0.1427, respectively. The lower variation indicates more consistent performance across the evaluated speaker classes.

6.3. Training and Validation Convergence

The accuracy and loss graphs of the proposed model for the NISP data set over ten epochs are shown in Figure 4. Plotting accuracy versus epoch is done in the left graph, and plotting loss against epoch is done in the right graph. The validation accuracy quickly rises from 0.17 at epoch 0 to 0.92 at epoch 9, as seen in Figure 4, demonstrating how quick and stable the learning process is.

Additionally, validation loss quickly decreases from a value of about 4.0 at epoch 0 to zero at epoch 9, indicating effective model convergence. At epoch 9, the training accuracy reaches a value of 0.50, growing at a somewhat slower rate than the validation accuracy. This is a multiple-class classification in a sample with 345 speakers, so this is not surprising.

6.4. Comparative Model Performance

Figure 5 displays a bar chart comparison of all tested models' Accuracy, Precision, Recall, and F1-Score. Across all four performance measures, the proposed model, which was trained for up to ten epochs, achieves the highest values among the evaluated models.

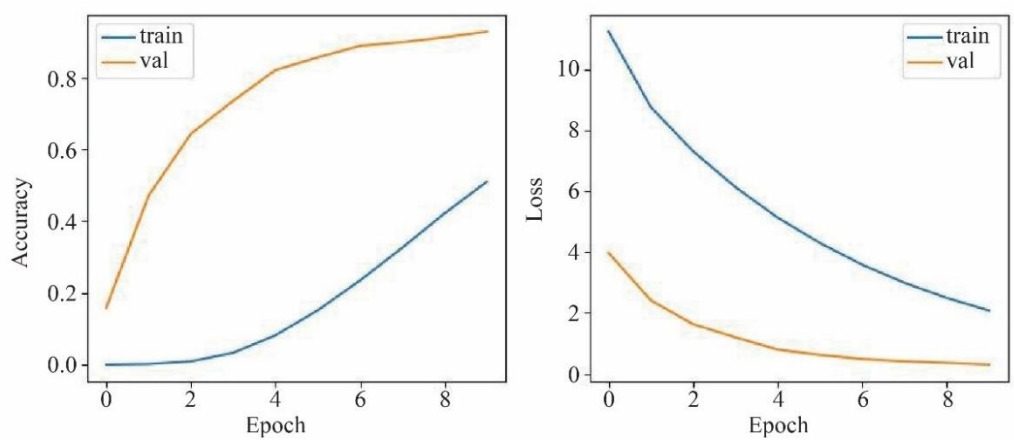

Fig. 4 Training and validation accuracy (left) and loss (right) curves of the proposed model over 10 epochs on the NISP dataset

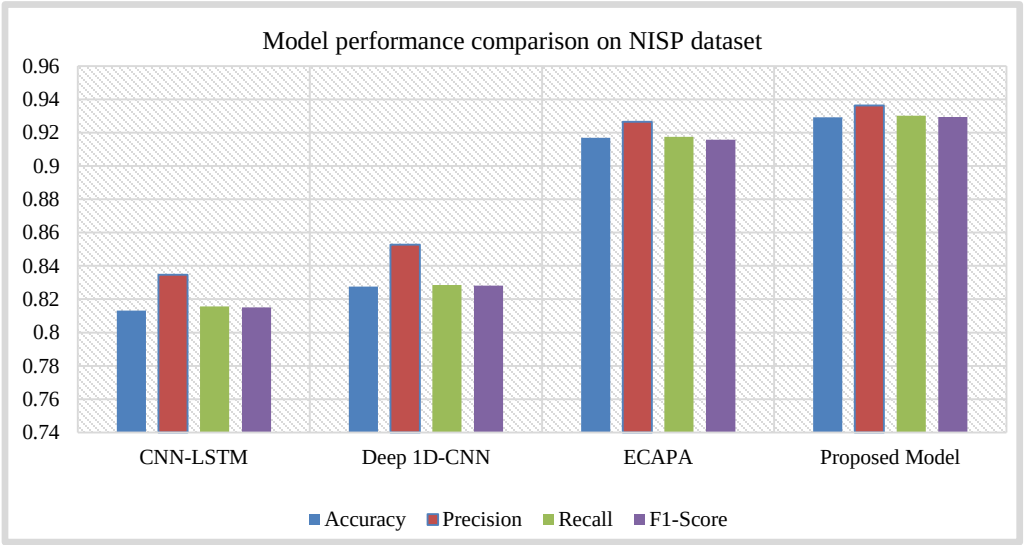

Fig. 5 Closed-set speaker identification performance comparison of the proposed ECAPA-TF model and baseline models in terms of accuracy, precision, recall, and F1-score on the NISP dataset

Figure 5 shows that the proposed ECAPA-TF model achieves the highest performance among the evaluated models. It attains an accuracy of 92.93% and a macro F1-score of 92.94%, compared with ECAPA (91.69% accuracy and 91.58% macro F1-score), Deep 1D-CNN (82.76% accuracy and 82.82% macro F1-score), and CNN-LSTM (81.32% accuracy and 81.50% macro F1-score).

6.5. Confusion Matrix Analysis

Row-normalized confusion matrices for each of the four models tested in NISP on a set of 345 speakers are displayed in Figures 6 through 9. True speaker labels are represented by the matrix's rows, whereas predicted speaker labels are represented by its columns. A cell entry along the main diagonal indicates a correct identification, while any value that is off-diagonal indicates an incorrect prediction.

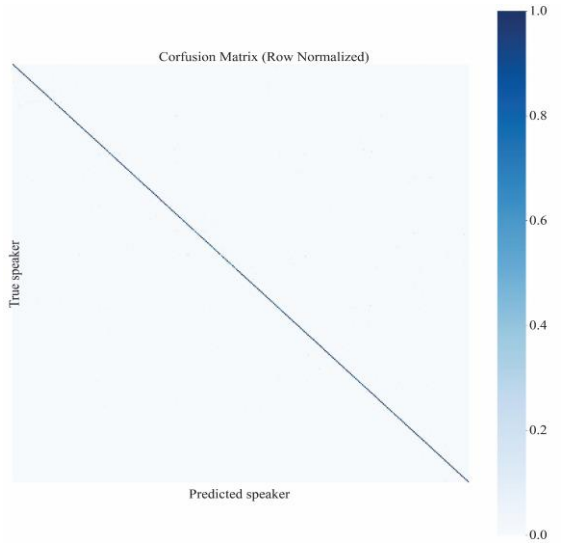

Fig. 6 Row-normalized confusion matrix - proposed model. Dense diagonal confirms very high speaker identification accuracy across all 345 speakers.

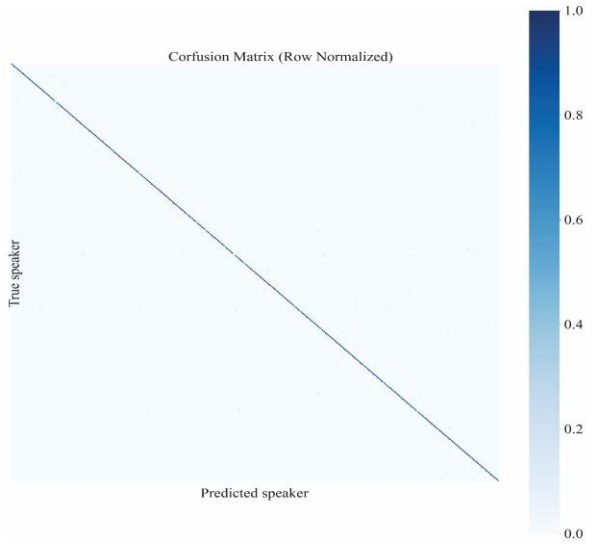

Fig. 7 Row-normalized confusion matrix-ECAPA only

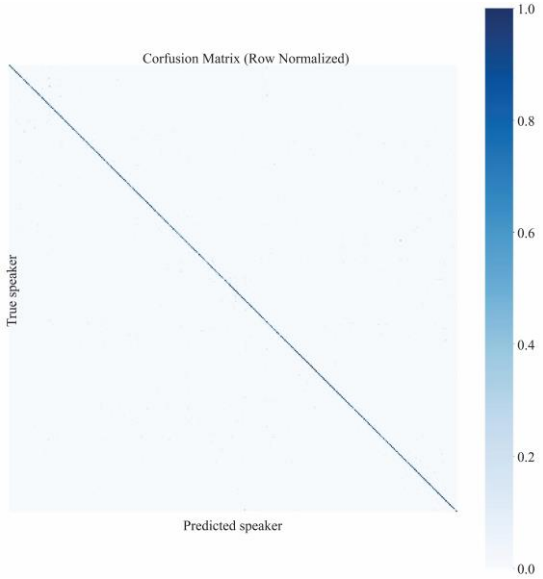

Fig. 8 Row-normalized confusion matrix - Deep 1D-CNN baseline. Moderate diagonal density with some visible off-diagonal scatter.

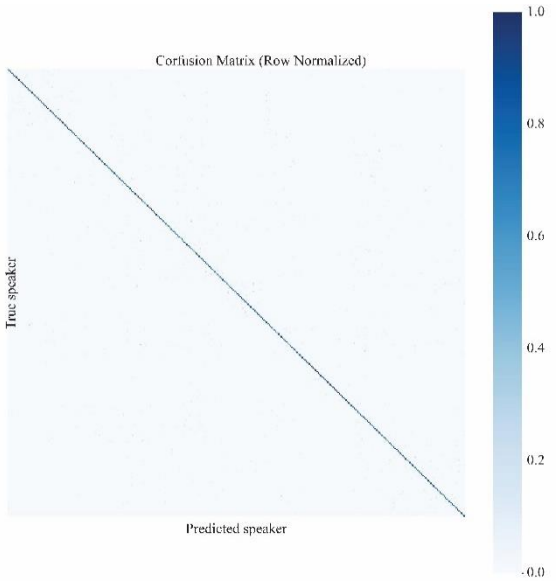

Fig. 9 Row-normalized confusion matrix - CNN-LSTM baseline. Similar pattern to Deep 1D-CNN with moderate classification accuracy.

The row-normalized confusion matrix of ECAPA-TF shows a dominant main diagonal and fewer off-diagonal entries than the CNN-LSTM and Deep 1D-CNN baselines.

This pattern is consistent with the macro recall of 93.02% and indicates lower inter-speaker misclassification under the evaluated test conditions.

For the row-normalized confusion matrix, the diagonal value for each speaker represents that speaker's recall. Therefore, the average of the diagonal values across all 345 speakers corresponds to the Macro Recall.

Table 6. Summary of confusion matrix performance indicators for all models on the NISP dataset

Model	Overall Accuracy (%)	Overall Misclassification Rate (%)	Macro Recall (%)	Mean Off-Diagonal Error (%)
Proposed model	92.93	7.07	93.02	6.98
ECAPA	91.69	8.31	91.76	8.24
Deep 1D-CNN	82.76	17.24	82.85	17.15
CNN-LSTM	81.32	18.68	81.57	18.43

6.6. ROC-AUC Analysis

The One-vs-Rest (OvR) approach is used to compute multiclass ROC-AUC because traditional ROC analysis is mathematically limited to binary (two-class) problems, whereas this project involves a complex multi-class environment with 345 distinct speaker classes. By breaking the multi-speaker classification task into 345 individual binary sub-problems (e.g., assessing "Speaker 1 vs. every other speaker", "Speaker 2 vs. every other speaker" and so on), OvR allows the system to compute the trade-offs between the True Positive Rate and False Positive Rate across all decision thresholds for each unique voice. The final ROC gives average performance across all speakers, helping in understanding how confidently the model distinguishes each speaker from the rest.

An assessment metric that gauges class-separability regardless of decision threshold is the Area Under the Curve (AUC). Macro AUC calculates the Area Under the Curve for each individual speaker independently and then takes the unweighted average of those scores. Micro AUC pools all voice samples from all speakers together into a single global pool, calculates the true positives and false positives globally, and then computes one final AUC. The Micro-Average and Macro-Average AUC values for each classifier are compared in Table 7.

Table 7. AUC-ROC scores (micro and macro average) for all models on the NISP dataset (One-Vs-Rest)

Model	Micro-Avg AUC	Macro-Avg AUC
Proposed Model	0.965	0.965
ECAPA	0.926	0.917
Deep 1D-CNN	0.914	0.914
CNN-LSTM	0.906	0.908

Since the evaluation involves 345 speaker classes, individual speaker-wise ROC curves are not displayed to avoid excessive visual clutter; macro-average AUC and micro-average AUC are used to provide an interpretable overall comparison.

The proposed model achieved micro-average and macro-average AUC values of 0.965. These were the highest values among the four evaluated models. Some speaker classes obtained an AUC of 1.000, indicating complete separation for those classes within the test set.

Table 7 shows that the Deep 1D-CNN model has an AUC score of 0.914, the CNN-LSTM model has an AUC score of 0.908, the ECAPA model has an AUC score of 0.917, and the proposed model is at the top of the list with an AUC score of 0.965.

6.7. Discussion

The experimental findings indicate that the improvement obtained by ECAPA-TF arises from the complementary modelling capabilities of its constituent stages. The ECAPA front-end captures local and multi-scale speaker-dependent acoustic characteristics, while the Transformer block extends the representation by modelling long-range temporal relationships across speech frames. The squeeze-excitation mechanism emphasizes informative channels, whereas attentive pooling provides an utterance-level representation by assigning greater importance to speaker-relevant frames. AAM-SoftMax further improves inter-speaker separation in the embedding space.

7. Conclusion

This study presented ECAPA-TF, a hybrid architecture for closed-set multilingual speaker identification using the NISP bilingual speech corpus. The framework combines ECAPA-style local and multi-scale acoustic modelling with Transformer-based long-range temporal modelling. Squeeze-excitation attention, attention pooling, and AAM-SoftMax were used to improve channel selection, utterance-level representation, and inter-speaker separation.

Under the common NISP experimental protocol, ECAPA-TF achieved higher accuracy, precision, recall, macro F1-score, and ROC-AUC than the internally implemented ECAPA, Deep 1D-CNN, and CNN-LSTM baselines.

Compared with the ECAPA baseline, the proposed model improved accuracy by 1.24 percentage points and macro F1-score by 1.36 percentage points. These comparisons are limited to models trained and evaluated under the same NISP setting and are not intended as direct numerical comparisons with published SOTA results obtained on different datasets.

From the analysis of the confusion matrix, the density of the diagonal elements shows that there is less confusion between speakers. Also, the lower standard deviation in speaker-level analysis, it shows the robustness of the model

against multilingual speech environments. Moreover, the research emphasizes the significance of integrating local modelling of speech signal characteristics and global learning of contextual information for enhancing the performance of multilingual speaker recognition tasks, especially in bilingual Indian speech scenarios where language changes and accent variations add another layer of complexity. The reported experiments were limited to a closed-set protocol using fixed-length MFCC sequences without online augmentation. Future work may examine language-disjoint evaluation, noisy speech, unseen speakers, self-supervised representations, and data augmentation.

References

[1] Juraj Kacur and Peter Truchly, "Acoustic and Auxiliary Speech Features for Speaker Identification System," *2015 57th International Symposium ELMAR (ELMAR)*, Zadar, Croatia, pp. 109-112, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[2] Nourah M. Almarshady, Adal A. Alashban, and Yousef A. Alotaibi, "Analysis and Investigation of Speaker Identification Problems Using Deep Learning Networks and the YOHO English Speech Dataset," *Applied Sciences*, vol. 13, no. 17, pp. 1-19, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[3] John H.L. Hansen and Taufiq Hasan, "Speaker Recognition by Machines and Humans: A Tutorial Review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74-99, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[4] Andreas Nautsch et al., "Preserving Privacy in Speaker and Speech Characterisation," *Computer Speech & Language*, vol. 58, pp. 441-480, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[5] Ajan Ahmed and Masudul H. Imtiaz, "Quantifying the Relationship Between Speech Quality Metrics and Biometric Speaker Recognition Performance Under Acoustic Degradation," *Signals*, vol. 7, no. 1, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[6] Ke-Ming Lyu, Ren-yuan Lyu, and Hsien-Tsung Chang, "Real-Time Multilingual Speech Recognition and Speaker Diarization System based on Whisper Segmentation," *PeerJ Computer Science*, vol. 10, pp. 1-19, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[7] S. Pruzansky and M.V. Mathews, "Talker-Recognition Procedure Based on Analysis of Variance," *The Journal of the Acoustical Society of America*, vol. 35, no. 11, pp. 2041-2047, 1964. [CrossRef] [Google Scholar] [Publisher Link]

[8] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 2, pp. 254-272, 1981. [CrossRef] [Google Scholar] [Publisher Link]

[9] S. Davis, and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357-366, 1980. [CrossRef] [Google Scholar] [Publisher Link]

[10] Hynek Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738-1752, 1990. [CrossRef] [Google Scholar] [Publisher Link]

[11] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19-41, 2000. [CrossRef] [Google Scholar] [Publisher Link]

[12] Kunal Thakur and Ramesh K. Bhukya, "Speaker Authentication Using GMM-UBM," *2022 IEEE 9th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, Prayagraj, India, pp. 1-6, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[13] Mitchell McLaren, and David van Leeuwen, "Improved Speaker Recognition when Using I-Vectors from Multiple Speech Sources," *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, pp. 5460-5463, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[14] Patrick Kenny et al., "Joint Factor Analysis versus Eigenchannels in Speaker Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435-1447, 2007. [CrossRef] [Google Scholar] [Publisher Link]

[15] Ehsan Variani et al., "Deep Neural Networks For Small Footprint Text-Dependent Speaker Verification," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, pp. 4052-4056, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[16] David Snyder et al., "X-Vectors: Robust DNN Embeddings for Speaker Recognition," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, pp. 5329-5333, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[17] David Snyder et al., "Deep Neural Network-Based Speaker Embeddings for End-to-End Speaker Verification," *2016 IEEE Spoken Language Technology Workshop (SLT)*, San Diego, CA, USA, pp. 165-170, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[18] Rahul Vijaykumar et al., "Descriptor: Extended-Length Audio Dataset for Synthetic Voice Detection and Speaker Recognition (ELAD-SVDSR)," *IEEE Data Descriptions*, vol. 3, pp. 93-99, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[19] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN-based Speaker Verification," *arXiv preprint*, pp. 3830-3834, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[20] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "VoxCeleb : A Large-Scale Speaker Identification Dataset," *arXiv preprint*, pp. 1-6, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[21] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "VoxCeleb2 : Deep Speaker Recognition," *arXiv preprint*, pp. 1-6, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[22] Vassil Panayotov et al., "Librispeech: An ASR Corpus Based on Public Domain Audio Books," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, Australia, pp. 5206-5210, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[23] Alexei Baevski et al., "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 12449-12460, 2020. [Google Scholar] [Publisher Link]

[24] Wei-Ning Hsu et al., "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451-3460, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[25] Zhanibek Kozhirbayev et al., "Speaker Recognition for Robotic Control via an IoT Device," *2018 World Automation Congress (WAC)*, Stevenson, WA, USA, pp. 1-5, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[26] Ayu Mawadda Warohma, Hilwadi Hindersah, and Dessi Puji Lestari, "Speaker Recognition Using MobileNetV3 for Voice-Based Robot Navigation," *2024 11th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*, Singapore, pp. 1-6, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[27] Jee-weon Jung et al., "RawNet: Advanced End-to-End Deep Neural Network using Raw Waveforms for Text-Independent Speaker Verification," *arXiv preprint*, pp. 1-5, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[28] Jee-weon Jung et al., *Pushing the Limits of Raw Waveform Speaker Recognition*, *arXiv preprint*, pp. 1-5, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[29] Seo-Hyun Kim, Tae-Wan Kim, and Keun-Chang Kwak, "Speaker Recognition Based on the Combination of SincNet and Neuro-Fuzzy for Intelligent Home Service Robots," *Electronics*, vol. 14, no. 18, pp. 1-22, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[30] Anuj Diwan et al., "Multilingual and Code-Switching ASR Challenges for Low-Resource Indian Languages," *arXiv preprint*, pp. 1-6, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[31] Mahesh K. Singh et al., "Feature Extraction and Classification Technique Based Speaker Identification System of Indian Regional Accent," *Franklin Open*, vol. 16, pp. 1-9, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[32] Kristiawan Nugroho et al., "Enhanced Indonesian Ethnic Speaker Recognition using Data Augmentation Deep Neural Network," *Journal of King Saud University Computer and Information Sciences*, vol. 34, no. 7, pp. 4375-4384, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[33] Vajratiya Vajrobol et al., "Enhancing Speaker Identification in Low-Resource Multilingual Languages using Hybrid MFCC-Chroma STFT and Transformer Encoder," *Multimedia Tools and Applications*, vol. 84, no. 35, pp. 44251-44285, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[34] Braveenan Sritharan and Uthayasanker Thayasivam, "Advancing Multilingual Speaker Identification and Verification for Indo-Aryan and Dravidian Languages," *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pp. 67-73, 2025. [Google Scholar] [Publisher Link]

[35] Ruijie Tao et al., "Self-Supervised Speaker Recognition with Loss-Gated Learning," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, pp. 6142-6146, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[36] Wenzao Li et al., "TDNN Architecture with Efficient Channel Attention and Improved Residual Blocks for Accurate Speaker Recognition," *Scientific Reports*, vol. 15, no. 1, pp. 1-13, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[37] Rafizah Mohd Hanifa, Khalid Isa, and Shamsul Mohamad, "Speaker Ethnic Identification for Continuous Speech in Malay Language Using Pitch and MFCC," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 19, no. 1, pp. 207-214, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[38] Muzamil Ahmed et al., "An Enhanced Deep Learning Approach for Speaker Diarization using TitaNet, MarbelNet and Time Delay Network," *Scientific Reports*, vol. 15, pp. 1-15, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[39] Md. Iftekharul Alam Efat et al., "Identifying Optimised Speaker Identification Model using Hybrid GRU-CNN Feature Extraction Technique," *International Journal of Computational Vision and Robotics*, vol. 12, no. 6, pp. 662-685, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[40] Weijun Pan et al., "The Speaker Identification Model for Air-Ground Communication Based on a Parallel Branch Architecture," *Applied Sciences*, vol. 15, no. 6, pp. 1-21, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[41] Hossein Zeinali et al., "BUT System Description to VoxCeleb Speaker Recognition Challenge 2019," *arXiv preprint*, pp. 4-7, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[42] Arsha Nagrani et al., "VOxSRC 2020: The Second VoxCeleb Speaker Recognition Challenge," *arXiv preprint*, pp. 1-8, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[43] A. Gusev et al., "SdSVC Challenge 2021 : Tips and Tricks to Boost the Short-duration Speaker Verification System Performance," *Proceedings of Interspeech 2021*, pp. 2307-2311, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[44] Andrew Brown et al., "VoxSRC 2021 : The Third VoxCeleb Speaker Recognition Challenge," *arXiv preprint*, pp. 1-8, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[45] Shareef Babu Kalluri et al., "NISP: A Multi-lingual Multi-accent Dataset for Speaker Profiling," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, pp. 6953-6957, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[46] Ashish Vaswani et al., "Attention is all you Need," *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Google Scholar] [Publisher Link]