Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Toward Deep Learning ECAPA-TDNN Model Enhancement for Speaker Recognition

Freha Mezzoudj[1] Chahreddine Medjahed[2], Ahmed Slimani[3]
[1]*National Polytechnic School of Oran, Oran, Algeria, freha.mezzoudj@enp-oran.dz*
[2]*Computer Science Department, Hassiba Benbouali Chlef University,Chlef, Algeria, c.medjahed@univ-chlef.dz*
[3]*LabRI-SBA Lab, Ahmed Draia University, ADRAR, Algeria, ah.slimani@esi-sba.dz*

ARTICLE INFO

Received: 18 Dec 2024

Revised: 10 Feb 2025

Accepted: 28 Feb 2025

ABSTRACT

The goal of artificial intelligence (AI) is to build intelligent machines or models that are able to learn, reason, solve problems, comprehend language and recognize patterns. A biometric system uses persons' physiological or behavioural features to recognize them. Applications like smartphones, border control, banking, and workplace access all make extensive use of these systems for identity management, security, and authentication. An essential component of the majority of biometric systems is a uni-biometric individual recognition system. With an emphasis on the voice, we suggest automatic person recognition systems that combine Deep Learning (DL) and Machine Learning (ML) approaches to ensure simplicity and efficiency. To accomplish this objective, we propose two strategies. First, we customized ECAPA-TDNN, a pretrained acoustic deep neural network, for individual speech recognition using transfer-learning technique. Second, we used transfer learning shaped ECAPA as a feature extractor for speech signals hybridized to a branch of ML algorithms as classifiers. We performed the training and testing of the systems using spoken acoustic signals gathered in real-world environments. The classification results indicate that the proposed methods make an interesting rate of accuracy. The overall accuracy was 100% in frame level with couple of hybridized models based on DL-ML models. The architecture based on DL feature extractor-ML classifier established in this study provided a foundation for promising behaviour biometric systems.

Keywords: artificial intelligence; machine learning; deep learning; ECAPA-TDNN; transfer Learning; biometry system; individual recognition; speech.

INTRODUCTION

In general, Artificial intelligence (AI) is based on the automated learning; models should be fed a significant amount of data pertaining to either a general or a particular task. In accordance with the intended tasks, those trained models are evaluated using unseen data to confirm their capacity to identify novel circumstances. Machine learning and Deep learning techniques are the foundations of the two important types of intelligent models. Deep learning has recently significantly advanced the disciplines of speaker and speech recognition. Various neural network-based architectures are proposed to improve the performance of speech recognition [1, 2], speech synthesis [3], speaker recognition systems [4] and many other biometric systems based on different traits or profiles.

In order to function properly, autonomous biometric systems are founded on the fundamental assumption that human characteristics are discriminatory. In general, fingerprints, iris patterns, facial features, voice, writing style and other characteristics or traits are examples of the distinctive qualities that are unique to each individual human being [5-9]. The data related to those traits can be used to construct uni-modal or multi-modal systems. Speaker recognition systems based on deep learning are becoming increasingly popular in both academic research and industrial applications [2].

In many applications, it is essential to have a speaker recognition module, which is based on biometric authentication system that uses a person's voice. Capturing distinctive voice features that set one speaker apart from another is its essential basic idea. Earlier systems employed statistical methods like i-vector representations [10], Mel-Frequency Cepstral Coefficients (MFCCs), and Gaussian Mixture Models (GMMs) [11]. Despite their respectable performance,

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

these models may faltered in real-world scenarios with difficult conditions. They are also difficult to develop and maintain since they rely on a variety of elements, such as training algorithms, acoustic models, language models, and vocabulary [12].

Deep neural networks eliminate the need for human feature extraction by automatically learning relevant features from spectrograms or raw audio. With the emergence of deep learning, several neural networks were proposed and used such as TDNN [13], x-vectors [14], YamNet [4] and ECAPA-TDNN [15]. In this paper, we concentrate on the latter model, which is called "Emphasized Channel Attention, Propagation and Aggregation- Time Delay Neural Network" (ECAPA-TDNN), an enhanced version of the TDNN and x-vectors. This is a pretrained audio deep learning classification model for speaker verification enhanced performance, incorporating the ideas from computer vision and speech processing, using both the residual Neural Network ResNet [8] and the TDNN model respectively.

ECAPA-TDNN is able for classifying a wide range of sounds. In [16], the ECAPA-TDNN model turned out to provide "who spoke when" in an audio, under both close-talking and distant-talking conditions. The results indicate that the proposed model achieve a good performance in the speaker diarization task. In [17], the authors propose a channel fusion technique that divides the spectrogram across the feature channels. Then, they add branches and increase the depth of the ECAPA-TDNN model. The obtained results using the enhanced were encouraging. In [18], the authors propose Branch-ECAPA-TDNN, which uses two parallel branches to extract features with n in both the global range and various local ranges, using multi-head self-attention and SE-Res2Block module respectively. Experimentations show that the proposed Branch-ECAPA-TDNN achieves good results on the VoxCeleb and CN-Celeb datasets. In clinical context, the authors in [19] use ECAPA-TDNN to develop a system for the depression speech detection in speech. The author trained the model with a corpus of 131 participants. In a close context, the authors of [20] used both speaker and contextual embedding vectors extracted from the ECAPA-TDNN and Wav2Vec2.0 respectively, to detect Stuttering in speech.

In this work, we explore the deep learning (DL) system ECAPA-TDNN as a baseline solution with transfer learning technique for speaker recognition task. Then, we apply a pipeline using the ECAPA-TDNN as deep feature extractor and feed his results to each one of four machine learning (ML) classifiers including Support Vector machines (SVM) [21], Random Forest (RF), K-nearest neighbours (K-NN) and Naïve Bayes (NB) [22]. A hybridization of both of these techniques can improve accuracy and speed of the intelligent biometric systems. The principal contributions in this work are as follow:

- Analyse the performance of the baseline model ECAPA-TDNN using signal (Speech).

- Propose architectural improvements, using the hybridization of four machine-learning algorithms (ML) as classifier to enhance recognition accuracy or speed of biometric systems based on ECAPA-TDNN.

- Validate the baselines and the obtained models in terms of accuracy and time processing.

This paper is composed of four sections. Section 2 introduces the proposed methods and the background about the DL and ML models used in those approaches and gives details of the used acoustic dataset. In Section 3, we describe and discuss the obtained results. Section 4 concludes with a brief overview of our findings.

METHODOLOGY

First, we introduce the overview of the key AI models that form the foundation of this work. This section includes also the essential details about the used dataset for training our models. Those details are essential to ease the follow the context and motivations behind the proposed models.

Deep learning Models

ECAPA-TDNN model

Inspired from human biological neural network, the artificial neural networks are models able to learn complex nonlinear relations that exist with inputs and outputs. The ECAPA-TDNN neural network architecture is used for speaker verification, which is the biometric task to recognise speakers from their voices. The model's architecture, as

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

originally described in [15], and Figure 1 illustrates it. To understand the whole architecture of ECAPA-TDNN, it is essential to distinguish between TDNN [13], x-vectors [14] and Squeeze-and-Excitation (SE) [23] and Res2Net [24, 8] blocks.

The standard Time-Delay Neural Network (TDNN) [13] is widely used in speech recognition software for the acoustic model, which converts the acoustic signal into a phonetic representation. The x-vector [14] architecture is a TDNN that applies statistics pooling to project variable-length utterances into fixed-length speaker characterizing embeddings. The x-vector architecture begins with audio inputs such as MFCCs or log Mel filter-bank energies, which pass through several TDNN layers designed to capture short-term speech patterns. A pooling layer then aggregates variable and length frames into a fixed-length vector by computing mean and variance across time. Finally, fully connected layers extract high-level speaker characteristics. The standard model uses 24-dimensional features and produces a 512-dimensional embedding, with about 4.2 million parameters [13, 14].

The original x-vector architecture is expanded upon by ECAPA-TDNN, which places greater emphasis on channel attention, propagation, and aggregation. Even as illustrated in figure 1, ECAPA-TDNN model extends this framework with key enhancements. TDNN layers are replaced by Res2Net [24] blocks with skip connections to capture multi-scale features. Squeeze-and-Excitation (SE) [23] blocks adapt channel weights, emphasizing critical speaker cues, while an attention-based pooling layer focuses on the most relevant frames. Inputs are 80-dimensional features, and the system outputs 192-dimensional embeddings. The configuration with 1024 convolutional filters includes SE-Res2Blocks with dilations and has approximately 14.7 million parameters. At end, “Multi-layer Feature Aggregation” combines complementary information prior to statistics pooling. The used AAM-softmax is better than the regular softmax loss in the context of fine-grained classification and verification problems [15].

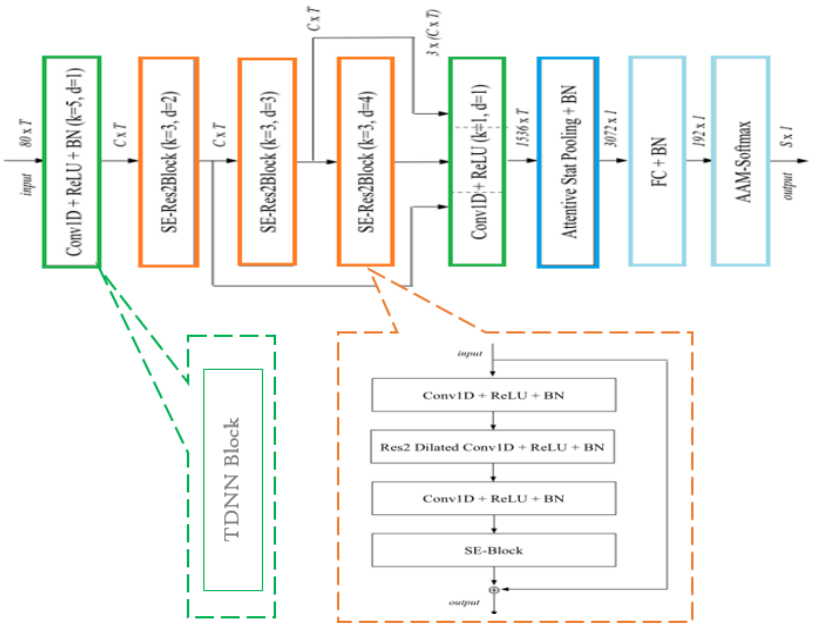

Figure 1: ECAPA-TDNN architecture inspired from [15].

For the ECAPA-TDNN model architecture, *k* represents the kernel size and *d* for dilation spacing (SE-Res2Blocks or *Conv1D* layers). T and *C* indicate the temporal and the channel dimension of the feature maps respectively. *S* denotes the training speakers’ number. For the SE-Res2Block architecture: the standard Conv1D layers have a kernel size of 1. Using kernel size k and dilation spacing d, the core Res2Net Conv1D with scale dimension s = 8 broadens the temporal context.

Machine learning Models

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Naïve Bayes

The naive Bayes algorithm (NB) is a machine supervised learning classifier. To complete a classification task, the algorithm uses likelihoods and prior probabilities of the features of the available data to estimate the posterior probability for each class and choose the one with the highest value. The naive Bayes model's primary premise is that the input variable distributions are conditionally independent [22].

K nearest neighbours

The K- nearest neighbours algorithm (K-NN) initially determines each data point's neighbourhood by identifying its K nearest neighbours or by locating every point inside a sphere with a specified radius, as illustrated with a simple manner in figure 2. According to their Euclidean distance, all nearby points are connected and labelled [22].

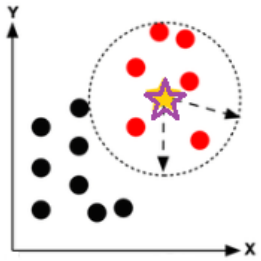

Figure 2: Principal's illustration of K-NN algorithm.

Support vector machines

Vapnik and his colleagues first suggested the Support Vector Machine (SVM) for binary classification. Later, they extended it for multi-classification tasks by employing techniques such as one-vs-rest or one-vs-one. SVMs are designed to find the best hyperplane for a bi-classification task that maximizes the margin, as illustrated in figure 3. With uses in speech recognition, image identification, text categorization, and other areas. According to the literature, SVM is a crucial machine learning technique [21, 12].

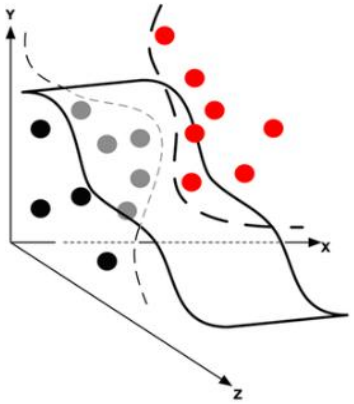

Figure 3: Principal's illustration of SVM.

Random Forest

A Random Forest (RF) algorithm combines many decision trees to enhance forecast accuracy and robustness. At each split, RF builds a set of trees that have been trained on various bootstrap samples and chooses subsets of features at random, as illustrated in figure 4. The RF averages the predictions made by each tree to aggregate the results for regression tasks [22].

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

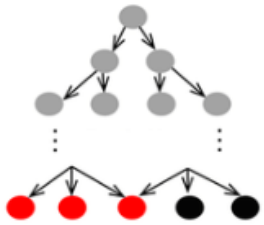

Figure 4: Principal's illustration of RF classifier.

Dataset

The VoxCeleb1 dataset, a comprehensive audio-visual collection of approximately 352 hours of speech that covers a variety of scenarios like interviews, news, and discussions with 1251 speakers (men and women), served as the dataset utilized to pre-train the model. More than 100,000 English sentences are included. YouTube videos are the source of the audio recordings. Because each extract is marked with a distinct speaker ID, tasks like voice identification and verification are made possible [25]. The dataset is split as following: 80% for training the models and 20% for their testing.

Evaluation Metric

We measure the speaker identification of the systems using the recognition accuracy, it represents the rate of speakers correctly identified by the number of total speakers, calculated as shown in Equation (1):

$$Accuracy = \frac{N_{correct}}{N_{total}} \quad (1)$$

Methodology of the Proposed Approach

According to the proposed approaches, we first explore and implement the deep learning (DL) system baseline solution based on ECAPA-TDNN with transfer learning technique. Second, we considere a pipeline using the ECAPA-TDNN as feature selector and we hybrid each one of four machine learning (ML) classifiers including Naïve Bayes (NB), K-nearest neighbours (K-NN), Support Vector machines (SVM) and Random Forest (RF). Figure 5 gives an illustration of the proposed systems' architecture.

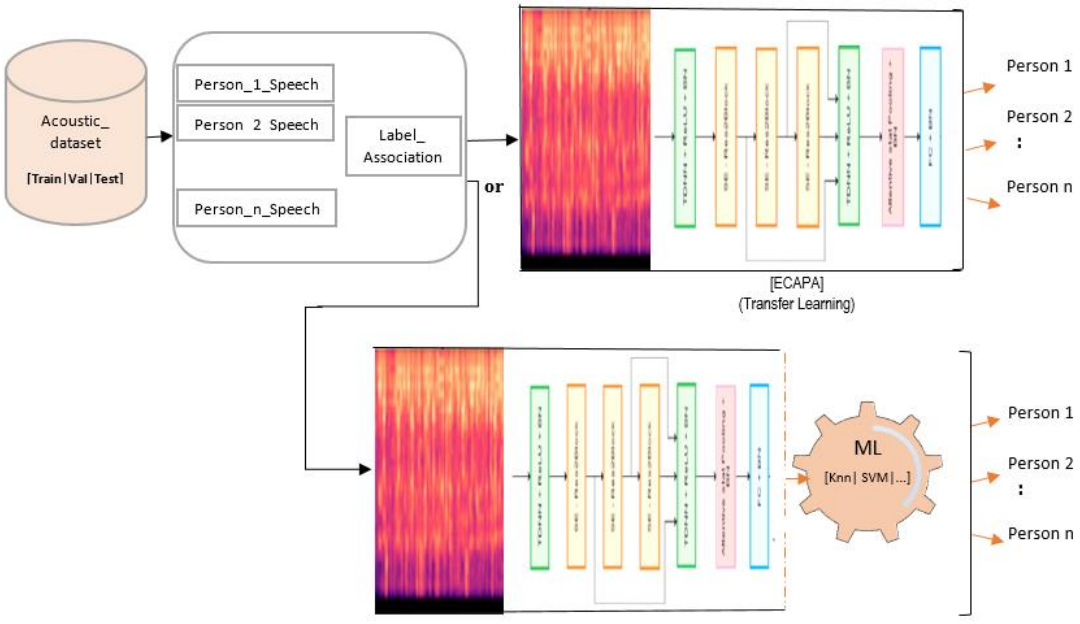

Figure 5: An illustration of the proposed biometric System based ECAPA-TDNN.

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

RESULTS

This section presents the proposed models' results according to the accuracy and the time processing. We will start with checking the baseline performance and then point out the improvements that we made with our hybridization based on ML algorithms. To determine a performance baseline, we used ECAPA-TDNN with transfer learning according to the number of speakers in the acoustic dataset, for the baseline unimodal biometric system. This pre-trained model, based on the ResNet and TDNN architectures, allows efficient extraction of audio features. Speakers are then classified using the obtained features.

This method offers good speech recognition performance in a realistic environment. Applied to a subset of 2,000 voice clips from the VoxCeleb1 dataset, it achieved a classification rate of 100%, from the 30th iteration, demonstrating a good efficiency. Table 1 summarizes the results obtained in terms of classification rates according to different epoch values:

#Epoch	Accuracy [%]	Time [s]
10	98.60	0.52
20	99.75	0.59
30	**100**	**1.20**
40	100	1.44
50	100	1.55
60	100	2.30

Table 1: Performance of Speaker recognition system based on ECAPA-tdnn.

However, for obtaining more targeted tasks such as fine-grained speaker recognition in less time, we apply the DL-ML hybridization; the results show that the choice of classification algorithm has a significant impact on speed system performance. The machine learning algorithms: Naïve Bayes (NB) algorithm, Support Vector Machine (SVM), K-NN algorithm (with cosine distance) and Random Forest (RF) achieve the same classification accuracy rate with 100%.

However, the use of ML models as classifiers provide the classification task during less time than the baseline ECAPA-TDNN alone. The fastest model is ECAPA-TDNN hybridized to Naive Bayes (NB). Although k-NN is quick to train, it takes longer to infer the acoustic datasets. Support Vector Machine (SVM) is slower than NB and occasionally on par with RF. Depending on the number of the generated trees; Random Forest (RF) is comparatively slower than NB and SVM. The Table 2 presents a comparison of speaker identification performance with various classifiers and ECAPA-TDNN embeddings. Processing time is equivalent to the average training/inference time each run and accuracy is provided for 30 and 60 iterations.

Model	Accuracy [%]	Time [s] 30-60 Iterations
ECAPA_TDNN (+ Transfer Learning)	100	**1.20** / 2.30
ECAPA_TDNN + NB	100	**0.90** / 1.80
ECAPA_TDNN + SVM	100	**1.10** /2.05
ECAPA_TDNN +RF	100	1.25 / 2.50
ECAPA_TDNN +k-NN	100	**0.95** / 1.90

Table 2: Performance of Speaker recognition system based on ECAPA-TDNN (DL) and Hybrid Systems (DL +ML).

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

CONCLUSION

The biometric systems should be able to distinguish one person from another, with sufficient level of accuracy in real time scenarios. We propose automatic person recognition systems using two strategies based on both Deep Learning (DL) and hybridization of DL & Machine Learning (ML) techniques focusing on the voice features. We first adapted ECAPA-TDNN, the acoustic deep neural network, for speaker recognition using transfer-learning technique. Second, we used transfer learning on ECAPA-TDNN as a feature extractor for speech features. Those embeddings are used as inputs to a couple of ML algorithms as classifiers: Naïve Bayes (NB) algorithm, Support Vector Machine (SVM), K-NN algorithm (with cosine distance) and Random Forest (RF).

The experimental findings validate the great discriminative power of the learnt acoustic features by showing that all hybrid models combining ECAPA-TDNN embeddings with traditional machine learning classifiers (Naive Bayes, SVM, Random Forest, and k-NN) obtained 100% accuracy. The primary difference in processing time was found between Random Forest and SVM, which took a little more time and Naive Bayes and k-NN, which offered the fastest classification. Overall, the results show that even basic classifiers can achieve flawless recognition provided strong speaker embeddings are extracted using ECAPA-TDNN. This means that efficiency and computational restrictions, rather than accuracy, are more important factors when choosing a backend model.

In this study, the results show good recognition results are achieved of 100% by the ECAPA-TDNN and all the ECAPA-TDNN- ML hybrid models. The important finding was related to the computing efficiency. As future works, we plan to explore other deeper transformer architectures and other machine learning algorithms, particularly for individual recognition in noisy environments.

REFRENCES

[1] Mezzoudj, F., & Benyettou, A. (2018). An empirical study of statistical language models: n-gram language models vs. neural network language models. International Journal of Innovative Computing and Applications, 9(4), 189-202.
[2] Padi, S., Sadjadi, S. O. , Sriram, R. D., Manocha, D. Improved speech emotion recognition using transfer learning and spectrogram augmentation, in: Proceedings of the 2021 International Conference on Multimodal Interaction, 2021, pp. 645–652
[3] Mezzoudj, F., Slimani, A., Chareddine, M., & wafa Krolkral, N. (2024, November). Experimental Study on Speech Synthesis Using Advanced Neural Network Models. In 2024 International Conference of the African Federation of Operational Research Societies (AFROS) (pp. 1-5). IEEE.
[4] Chahreddine Medjahed, Freha Mezzoudj, Ahmed Slimani, Narimane Wafaa Krolkral. YAMNet Accuracy Enhancement for Speaker Recognition. JISEM, (pp. 753 – 759)
[5] Ahmed Slimani, Abdellatif Rahmoun, Chahreddine Medjahed, Freha Mezzoudj. Improving Fake Profile Detection: A Hybrid Machine Learning Approach with Negative and Clonal Selection. JISEM, (pp. 868 - 876). DOI: https://doi.org/10.52783/jisem.v10i56s.12030.
[6] Medjahed, C., Mezzoudj, F., Rahmoun, A., & Charrier, C. (2020, June). On an empirical study: face recognition using machine learning and deep learning techniques. In Proceedings of the 10th International Conference on Information Systems and Technologies (pp. 1-9).
[7] Medjahed, C., Rahmoun, A., Charrier, C., & Mezzoudj, F. (2022). A deep learning-based multimodal biometric system using score fusion. IAES Int. J. Artif. Intell, 11(1), 65.
[8] Mezzoudj, F., & Medjahed, C. (2024). Efficient masked face identification biometric systems based on ResNet and DarkNet convolutional neural networks. International Journal of Computational Vision and Robotics, 14(3), 284-303.
[9] Medjahed, C., Mezzoudj, F., Rahmoun, A., & Charrier, C. (2023). Identification based on feature fusion of multimodal biometrics and deep learning. International Journal of Biometrics, 15(3-4), 521-538.
[10] Dehak, N.; Kenny, P.J.; Dehak, R.; Dumouchel, P.; Ouellet, P. Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process. 2011, 19, 788-798.
[11] Rabiner, L. R., & Schafer, R. W. (2011). Theory and applications of digital speech processing. Pearson Education.

[12] Mezzoudj, F., & Benyettou, A. (2012). On the optimization of multiclass support vector machines dedicated to speech recognition. In Neural Information Processing: 19th International Conference, ICONIP 2012, Doha, Qatar, November 12-15, 2012, Proceedings, Part II 19 (pp. 1-8). Springer Berlin Heidelberg.
[13] Snyder, D., Garcia-Romero, D., & Povey, D. (2015, December). Time delay deep neural network-based universal background models for speaker recognition. In 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU) (pp. 92-97). IEEE.
[14] Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; Khudanpur, S. X-vectors: Robust DNN embeddings for speaker recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15-20 April 2018.
[15] Desplanques, B.; Thienpondt, J.; Demuynck, K. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In Proceedings of the Interspeech, 2020, Shanghai, China, 25-29 October 2020; pp. 3830-3834.
[16] Guo, J., Zhu, J., Lin, S., & Shi, F. (2024, March). ECAPA-TDNN Embeddings for Speaker Recognition. In 2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT) (pp. 1488-1491). IEEE.
[17] Zhao, Z., Li, Z., Wang, W., & Zhang, P. (2023, June). Pcf: Ecapa-tdnn with progressive channel fusion for speaker verification. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.
[18] Yao, J., Liang, C., Peng, Z., Zhang, B., & Zhang, X. L. (2023, August). Branch-ECAPA-TDNN: A parallel branch architecture to capture local and global features for speaker verification. In Proc. Interspeech (pp. 1943-1947).
[19] Wang, D., Ding, Y., Zhao, Q., Yang, P., Tan, S., & Li, Y. (2022, September). ECAPA-TDNN Based Depression Detection from Clinical Speech. In Interspeech (pp. 3333-3337).
[20] Zhou, J., Li, Y., & Yu, H. (2025). Infant Cry Emotion Recognition Using Improved ECAPA-TDNN with Multiscale Feature Fusion and Attention Enhancement. arXiv preprint arXiv:2506.18402.
[21] Cortes, C. and Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3):273-297.
[22] Bishop, C. M., & Nasrabadi, N. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: springer.
[23] [se] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation networks," in Proc. IEEE/CVF CVPR, 2018, pp. 7132–7141.
[24] [res] S. Gao, M.-M. Cheng, K. Zhao, X. Zhang, M.-H. Yang, and P. H. S. Torr, "Res2Net: A new multi-scale backbone architecture," IEEE TPAMI, 2019.
[25] Nagrani, A., Chung, J. S., & Zisserman, A. (2017). Voxceleb: a large-scale speaker identification dataset. arXiv preprint arXiv:1706.08612.