Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Toward Robust Deep Learning Models based on YAMNet vs ECAPA-TDNN for Speaker Recognition

Freha Mezzoudj[1,2] Chahreddine Medjahed[3], Ahmed Slimani[4,5], Ali Ould Krada[3]

[1]Department of Computer Software Engineering, National Polytechnic School of Oran- Maurice Audin, Oran 31000, Algeria
[2]LTE Laboratory, freha.mezzoudj@enp-oran.dz
[3]Computer Science Department, Hassiba Benbouali Chlef University, Chlef, Algeria, c.medjahed@univ-chlef.dz
[4]LabRI-SBA Lab, [5]Ahmed Draia University of Adrar, Adrar, Algeria, ah.slimani@esi-sba.dz
[3]Computer Science Department, Hassiba Benbouali Chlef University, Chlef, Algeria, a.ouldkrada@univ-chlef.dz

ARTICLE INFO

Received: 18 Dec 2024

Revised: 10 Feb 2025

Accepted: 28 Feb 2025

ABSTRACT

The importance of biometric identification and speaker recognition is growing in today's culture. Neural networks; especially the deep ones, are now frequently used to extract speaker attributes. Despite its limited ability to acquire fully comprehensive speech features, the YAMNet and ECAPA-TDNN models can get relevant context information by exploring acoustic feature parameters to pattern matching. However, in noisy environments, the background noise reduces speech quality and intelligibility, which make speaker identification challenging task. It is important to verify the biometric model's capacity for generalization and enable precise speaker recognition even in noisy environments. To assess the efficiency and the robustness of the introduced models in speaker identification and recognition, comparisons with YAMNet, ECAPA-TDNN and both of them hybridized with Machine Learning (ML) algorithms are conducted. The overall accuracies were affected by the noises in frame level, when using both the deep neural networks based on deep learning (DL) and the hybridized DL-ML models. The obtained results and the comparison presented in this paper furnish a founding for promising behaviour for robust biometric systems. The best results are obtained with ECAPA-TDNN model. In addition, the hybridization methods can perform good rate of accuracies especially when the DL models are hybridized with Support vector machines (SVM) in noisy environment.

Keywords: deep learning; artificial intelligence; machine learning; ECAPA-TDNN; YAMNet; biometry system; speaker; speech; noise

INTRODUCTION

With benefits over more conventional authentication techniques like ID cards or passwords, biometric technologies have emerged as crucial instruments for safe and trustworthy personal identity. Biometrics improve accuracy, convenience, and fraud resistance by assessing distinct physiological or behavioral traits, such as iris patterns, fingerprints, voice, or face features. Voice biometrics is unique among these modalities because to its non-intrusiveness, ease of acquisition, and applicability for hands-free or distant authentication.

Biometric systems consider the naïve premise that human features are discriminatory in the context of security. The voice is one of several characteristics that are specific to each human being, including fingerprints, iris, facial features, writing style, speech, etc. [1–9]. Building uni-modal or multi-modal systems requires the information pertaining of such characteristics. Deep learning-based speaker recognition systems are gaining popularity in both academic and commercial settings [10].

However, in noisy settings where background noise, reverberations, or communication channel distortions can deteriorate speech quality and lower recognition accuracy, voice-based biometric systems encounter considerable difficulties. In practical applications like contact centres, smart homes, and mobile authentication, these restrictions are very important. To improve the biometric systems' robustness on the noise, additive noise-data augmentation method are commonly used. Advanced noise reduction and speech enhancement algorithms, reliable feature extraction methods, and the use of machine learning and deep learning models for adaptive noise compensation are also some of the solutions that have been put forth to address these problems [11].

Speech applications involving audio analysis such as speech recognition, speaker identification, etc. increasingly use general-purpose embedding audio representations and transfer learning with audio pretrained networks, which helps our models to conduct a new task by transferring knowledge from another similar task. Deep learning based Models such as TDNN [12], x-vectors [13], YAMNet [14, 3], and ECAPA-TDNN [15, 4] can be used for this purpose. This research focuses on the two latter models, which are considered as good and recent versions of pretrained audio deep learning classification model for speaker verification [3, 4].

In [16], the authors compared the Time Delay Neural Network (TDNN) model with ResNet model in biometry field. According to their results, ResNet network yield better in individual recognition tasks with reverberation and noise than TDNN. Even, in [17], the authors explored speaker adaptation in different noise conditions using transformers and wav2vec 2.0. For the extraction of speaker embeddings, they used x-vector systems, ECAPA-TDNN, and i-vectors. The tests were done on two databases; LibriSpeech and Switchboard. The proposed method relied on concatenating speaker vectors with acoustic features to improve speech recognition systems. According to their experimentation, and when more noise is added to the input speech, the effect on transformer models is stronger. The results indicate that the use of ECAPA-TDNN and x-vectors embeddings helps to increase robustness in noisy environments and they outperform i-vectors as speaker representations. Additionally, the wav2vec 2.0-based system was weak under loud noise conditions.

In [18], the authors propose a partial additive-noise speech (PAS) method, which aims to train systems to be robust in noisy environments for Speaker Verification task. The experimental results show that PAS outperforms additive noise in terms of equal error rates (EER), with ameliorations in SE-ResNet34 and ECAPA-TDNN models. Their experiments was conducted on MUSAN and VoxCeleb1 datasets. In [19], the authors evaluate two embeddings – YAMNet, and OpenL3 on the monophonic UrbanSound8K and the polyphonic SONYC-UST urban datasets. They employ many distance measures such as Frechet Audio Distance (FAD), to estimate the effect of channel effects. In terms of the embedding performance, they get OpenL3 more robust than YAMNet.

To avoid the brusque failure of the machines in industry, it is important to detected rapidly any defects in gear. In [20], the authors use the YAMNet network for intelligent fault detection of gear pairs using transfer learning with noises and vibration data. Experimental setups were done with two similar gear pair sub-assemblies, one with gear tooth failure condition and another in healthy condition. Audio signals of gears was amassed using an electret microphone. The extracted features by YAMNet; from audio file's spectrograms, were used to classify gear noise as faulty or healthy. The results achieve 95 % prediction accuracy for random test data.

In this work, we performed all of our experiments using noise augmentation to increase the dataset's diversity and help the proposed models in adapting to noise setting, as the dataset we used is noise-free. We explore two deep learning (DL) systems YAMNet and ECAPA-TDNN as baseline solutions with transfer learning technique for speaker recognition task. Then, we apply a pipeline using both of them separately as deep feature extractor and feed his results to each one of four machine learning (ML) classifiers including Support Vector machines (SVM), Random Forest (RF), K-nearest neighbours (K-NN) and Naïve Bayes (NB) [21] with clean [3, 4] and noisy acoustic data. The important contributions in our work are as follow:

- Comparing the baseline performance of the both deep Neural Networks YAMNet and ECAPA-TDNN using clean and noisy acoustic data.

- Analyse the performance of the hybridized machine-learning algorithms (ML) to YAMNet and ECAPA-TDNN using clean and noisy acoustic data.

- Validate the results of the obtained models in term of accuracy.

-The use of two kind of noises.

The paper is composed from five section. Section 2 introduce a background about DL and ML models used in our proposed approaches, and detail of the used clean acoustic dataset and the additive noises. Section 3 describes and discuss the obtained results. Section 4 concludes with an overview of the important findings and perspectives.

BACKGROUND

In this section, we introduce an overview of AI models based on Deep Learning and Machine Learning that are explored in this work.

Deep learning Models

Inspired from human biological neural network, the artificial neural networks are models able to learn complex nonlinear relations between inputs and outputs. Both YAMNet and ECAPA-TDNN neural network architectures are proposed and used in the literature for speaker and speech recognition. The Mel Spectrogram extracted from raw audio signal in the time-frequency domain is useful in training those deep Neural Networks to distinguish between speech of different speakers. By reusing the weights in the pretrained network, the transfer learning technique is able to reduce the cost and the computing time of training new models.

YAMNet model

Google's YAMNet (Yet Another Multi-scale Convolutional Neural Network) [14, 22, 3] is a deep learning model based on MobileNetV1. YAMNet uses log mel spectrograms as input. The MobileNet v1 architecture, upon which YAMNet is based, was created for effective deep learning on mobile devices. Each convolution layer is divided into two smaller steps: one that filters each channel independently, and another that combines them. This is known as depthwise separable convolutions. This speeds up computation and lowers the number of parameters. Audio is fed into the model, which transforms it into a log-mel spectrogram—a 2D image-like representation of sound—and then runs it through a series of batch normalization and convolutional layers with ReLU activations. Following multiple feature extraction layers, YAMNet classifies the sound into one of 521 AudioSet categories using global average pooling then a fully connected layer with softmax. The standard architecture is illustrated in figure 1.

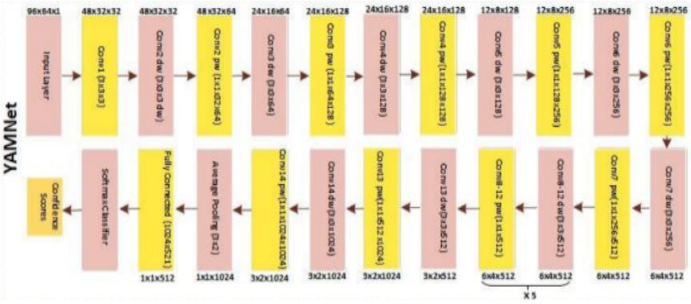


Figure 1: Illustration of the YAMNet architecture [14].

ECAPA-TDNN model

The architecture of ECAPA-TDNN model; as originally described in [15], is relative to TDNN, x-vectors, Squeeze-and-Excitation (SE) and Res2Net blocks' architectures. Even as illustrated in figure 2, ECAPA-TDNN model contains many key enhancements. To collect multi-scale features, Res2Net blocks with skip connections are used in place of TDNN layers. With an attention-based pooling layer that concentrates on the most pertinent frames, squeeze-and-excitation (SE) blocks adjust channel weights to highlight important speaker signals. The system produces 192-dimensional embeddings from inputs that are 80-dimensional characteristics. The 1024 convolutional filter configuration has over 14.7 million parameters and incorporates SE-Res2Blocks with dilations. Lastly, "Multi-layer Feature Aggregation" concatenates the final frame-level feature map with intermediate feature maps from previous layers to aggregate complementary information before statistics pooling. In the domain of fine-grained classification & verification problems, the AAM-softmax is superior to the standard softmax loss [15].

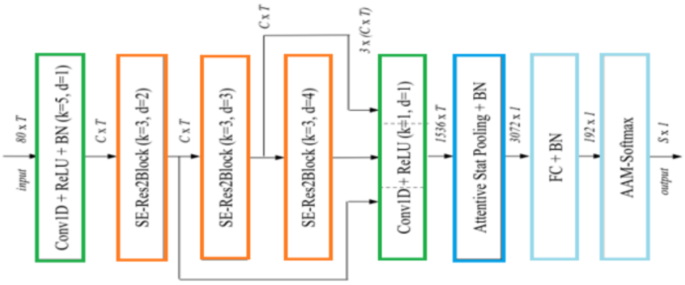


Figure 2: ECAPA-TDNN architecture inspired from [15, 4].

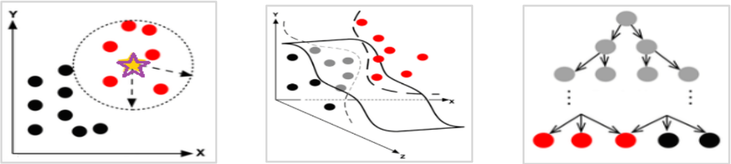
Machine learning Models

Naive Bayes algorithm (NB) is a supervised machine learning classifier. To complete a classification task, the algorithm uses likelihoods and prior probabilities of the features of the available data to compute the posterior probability for all classes and choose the one with the highest value. The naive Bayes model's primary premise is that the input variable distributions are conditionally independent [21].

K nearest neighbors algorithm (K-NN) initially determines each data point's neighborhood by identifying its K nearest neighbors or by locating every point inside a sphere with a specified radius, as illustrated with a simple manner in figure 3 (a). According to their Euclidean distance, all nearby points are connected and labeled [21].

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

Random Forest is a learning algorithm that combines many decision trees to improve forecast accuracy and sturdiness is the Random Forest (RF) regressor. 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 3 (c). The RF averages the predictions made by each tree to aggregate the results for regression tasks [21].



(a) (b) (c)

Figure 3: Principal's illustrations of ML algorithms for classification tasks: (a) K-nearest neighbors, (b) Support vector machines and (c) Random Forest.

EXPERIMENTATION SETTINGS

This section includes the essential details about the used dataset for training our models, the evaluation's metric used, the types of noise explored and so on.

Dataset

We use the VoxCeleb dataset for our experiments. The VoxCeleb1 dataset, a comprehensive audio-visual collection of approximately 352 hours of speech that covers a variety of YouTube's scenarios like interviews, news, and discussions with 1251 speakers (men and women). More than 100,000 English sentences are included. Each extract is marked with a distinct speaker ID to ease the tasks like voice verification and identification [23]. The dataset is splited as following: 80% for training and the remnant for models' testing.

Evaluation Metric

To compare the quality of our proposed model over other models, it is essential that many representative models trained under identical conditions are used. There are recognition errors in speaker recognition. False negative rate (FNR) is when the even group of speakers is wrongly identified as different individuals, and false positive rate (FPR) is when different speakers are incorrectly identified as the same individual. The equal error rate EER is the relative equilibrium point threshold when FNR and FPR are nearly equal. Therefore, we evaluate the speaker identification of our system; under balanced test conditions, using the recognition accuracy which is simply one minus the EER, which represents the rate of speakers correctly identified by the number of total speakers, calculated as shown in Equation (1):

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

Noisy data in experiments

We add a specific signal-to-noise ratio to the noise audio. The maximum Signal-to-Noise Ratio (SNR) is established at 50 dB, with a minimum SNR set to 10 decibels (dB). Two types of noise were exploited Gaussian and Cafeteria noises. We have take care to adjust SNR to control how strong the noise is relative to speech.

When noise is added at a high SNR ratio, the speech remains clear and the background is feebly audible. At a moderate ratio, the background chatter becomes almost as loud as the speaker, making it harder to distinguish the voice, while at a low or a negative ratio, the noise surpass the signal and the speech becomes challenging to recognize [24]. According to such as recommendation, we have choose our scenarios, presented in the future section.

Gaussian Noise

The Gaussian noise refers to random noise with a Gaussian (normal) distribution in terms of amplitude. It is also known as white Gaussian noise (WGN) when its power is constant across all frequencies. It is represented using equation 2:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (2)$$

Where, μ is the mean and σ^2 is the variance, which controls the noise power.

Cafeteria noise

Cafeteria noise alludes to the cacophonous atmosphere of a cafeteria, which includes general conversation with a lot of people conversing at once, cutlery noises, shifting chairs, and echoes. For testing voice recognition, speaker identification, or hearing aids, it is frequently regarded as a realistic and difficult scenarios with noise. MUSAN's cafeteria noise corpus [11] is used in the current context.

METHODOLOGY

Methodology of the Proposed Approach

According to the proposed approaches, we explore ten deep learning (DL) systems: YAMNet and ECAPA-TDNN both of them with transfer learning technique, YAMNet hybridized each time with Naïve Bayes (NB), K-nearest neighbours (K-NN), Support Vector machines (SVM) & Random Forest (RF), and ECAPA-TDNN hybridized each time also with the four ML algorithms. The data used for training those models are infected by Gaussian and Cafeteria noise in different scenarios. Noise data is preprocessed to extract Mel Spectrogram and is given at the input size requirement of the YAMNet and the ECAPA-TDNN pretrained networks and the hybridized models. The final connected layer is replaced, and hyperparameters are fine-tuned to improve accuracy for the biometric task. Figure 4 illustrates the global architecture of the proposed systems:

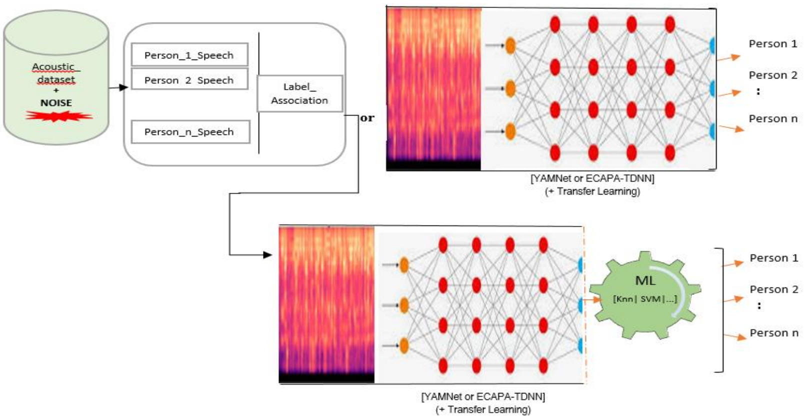


Figure 4: An illustration of the proposed biometric System based deep neural networks and deep neural networks hybridized with machine learning algorithms in a noisy environment.

RESULTS

This section presents the accuracies experimentations. We will start with checking the baseline performance and then point out the improvements that we made with our hybridization models based on ML algorithms. To determine a performance baseline, we used ECAPA-TDNN with transfer learning according to the speakers' number, the acoustic dataset, for the first unimodal biometric system. This pre-trained model, based on the ResNet and ECAPA-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.

Table1: The results obtained in terms of classification rates with models trained on clean data.

YAMNet Accuracy [%]					ECAPA-TDNN Accuracy [%]				
.	+ SVM	+ RF	+ KNN	+ NB	.	+ SVM	+ RF	+ KNN	+ NB
90.75	**92.75**	86.75	81.75	53.75	**100**	**100**	**100**	**100**	**100**

Table 2: the results obtained in terms of classification rates without and with infection of **Gaussian noise** [3, 4].

Noise Rate [%]	YAMNet Accuracy [%]		ECAPA-TDNN Accuracy[%]	
	.	+ SVM	.	+ SVM
0	**90.75**	**92.75**	**100**	**100**
10	80.25	83.00	99.95	99.97
20	72.75	78.50	99.77	99.80
30	68.75	76.75	99.50	99.55
40	67.25	74.75	99.25	99.35
50	64.00	71.25	98.00	98.15
60	63.00	70.00	97.00	97.25

Table 3: the results obtained in terms of classification rates without and with infection of **Cafeteria noise**.

Noise Rate [%]	**YAMNet Accuracy [%]**		**ECAPA.TDNN Accurac[%]**	
	.	**. SVM**	**.**	**.SVM**
0	**90.75**	**92.75**	**100**	**100**
10	73.50	87.50	**100**	100
20	68.25	81.00	99.90	100
30	62.00	71.50	99.75	100
40	56.25	67.75	99.50	99.90
50	48.50	63.75	99.25	99.75
60	47.00	62.00	98.70	99.25

In order to simulate challenging situations, we have use a noisy environment with both Gaussian then cafeteria noises. Table 1 summarizes the results obtained in terms of classification rates according to the best number of 60 epochs (when trying with 10, 20, 30, 40, 50), with models trained on clean data. However, table 2 summarizes the results obtained in terms of classification rates according to the best number of 60 epochs, without and with Gaussian noise injection. In addition, table 3 summarizes the results obtained in terms of classification rates according to the best number of 60 epochs, without and with infection of Cafeteria noise. The results show that the latter perturbs more the models. ECAPA-TDNN has contribute to fine-grained speaker recognition in clean and noisy situation better than the YAMNet. The application of the DL-ML hybridization has a significant impact on system performance especially with Support Vector Machine (SVM).

CONCLUSION

The biometric systems should be able to distinguish one person from another, with sufficient level of accuracy. Deep Learning (DL) and hybridization of DL & Machine Learning (ML) techniques; as classifiers: Naïve Bayes (NB) algorithm, K-NN algorithm (with cosine distance), Random Forest (RF) and Support Vector Machine (SVM), focusing on the voice features was explored. We propose in this paper to investigate the recognition and identification of individual in noisy situations. We adapted ECAPA-TDNN, the acoustic deep neural network, for speaker recognition using transfer-learning technique. We applied transfer learning on ECAPA-TDNN and extracted speech features. Those embeddings are used as inputs to a couple of ML algorithms pipelines.

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. We find that ECAPA-TDNN is more robust than YAMNet, with and without noise. In the future, we can further optimize our model and verify it on more datasets to improve the classification performances. Furthermore, dependability under unfavorable acoustic conditions can be further enhanced by integrating multi-modal biometrics, which combine voice with facial or lip movements.

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] 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.

[3] Medjahed, C. Mezzoudj, F. Slimani, A. Krolkral. N. W. YAMNet Accuracy Enhancement for Speaker Recognition, Journal of Information Systems Engineering and Management. Vol. 10 No. 56s (2025), pp. 462 – 470.

[4] Mezzoudj, F. Medjahed, C. Slimani, A. Toward Deep Learning ECAPA-TDNN Model Enhancement for Speaker Recognition. Journal of Information Systems Engineering and Management, Vol. 10 No. 59s (2025). pp. 377-384.

[5] Slimani, A. Rahmoun, A. Medjahed, C. Mezzoudj, F. 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] 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.

[11] Snyder, D. Chen, G. Povey, D. Musan: A music, speech, and noise corpus. ArXiv abs/1510.08484 (2015).

[12] 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.

[13] 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.

[14] Mahum, R., Irtaza, A., & Javed, A. (2023). EDL-Det: A robust TTS synthesis detector using VGG19-based YAMNet and ensemble learning block. IEEE Access, 11, 134701-134716.

[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] Mohammad Amini, M. Matrouf, D. Bonatsre, J.F. Dowerah, S. Serizel, R. Jouvet, D. (2022). A comprehensive exploration of noise robustness and noise compensation in resnet and tdnn-based speaker recognition systems in 2022 30th European Signal Processing Conference (EUSIPCO). pp. 364–368.

[17] Wagner, D., Baumann, I., Bayerl, S. P., Riedhammer, K., & Bocklet, T. (2023, December). Speaker adaptation for end-to-end speech recognition systems in noisy environments. In 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 1-6). IEEE

[18] Kim, W., Shin, H. S., Kim, J. H., Heo, J., Lim, C. Y., & Yu, H. J. (2023). Pas: Partial additive speech data augmentation method for noise robust speaker verification. arXiv preprint arXiv:2307.10628.

[19] Srivastava, S., Wu, H. H., Rulff, J., Fuentes, M., Cartwright, M., Silva, C., ... & Bello, J. P. (2022, August). A study on robustness to perturbations for representations of environmental sound. In 2022 30th European Signal Processing Conference (EUSIPCO) (pp. 125-129). IEEE.

[20] Atil, S., & Wani, K. (2023). Gear fault detection using noise analysis and machine learning algorithm with YAMNet pretrained network. Materials Today: Proceedings, 72, 1322-1327

[21] Bishop, C. M., & Nasrabadi, N. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: springer.

[22] Plakal. M. (Jan. 2020). Yet Another Mobile Network (YAMNet). [Online]. Available: https://github.com/tensorflow/models/tree/master/ research/audioset/YAMNet

[23] Nagrani, A., Chung, J. S., & Zisserman, A. (2017). Voxceleb: a large-scale speaker identification dataset. arXiv preprint arXiv:1706.08612.

[24] Lee, J. Y., Lee, J. T., Heo, H. J., Choi, C. H., Choi, S. H., & Lee, K. (2015). Speech recognition in real-life background noise by young and middle-aged adults with normal hearing. J. of audiology & otology, 19(1), 39.