Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

YAMNet Accuracy Enhancement for Speaker Recognition

Chahreddine Medjahed[1], Freha Mezzoudj[2], Ahmed Slimani[3], Narimane Wafaa Krolkral[4]
[1]Computer Science Department, Hassiba Benbouali Chlef University,c.medjahed@univ-chlef.dz
[2]The National Polytechnic School of Oran Algeria, freha.mezzoudj@enp-oran.dz
[3]LabRI-SBA Lab Ahmed Draia University – ADRAR, ah.slimani@esi-sba.dz
[4]EEDIS Laboratory, Djillali Liabes University,wafaa.krolkral@univ-sba.dz

ARTICLE INFO

Received: 28 Dec 2024

Revised: 18 Feb 2025

Accepted: 26 Feb 2025

ABSTRACT

A biometric system is able to identify or verify individuals based on their physiological traits or behavioural characteristics. These systems are widely used for security, authentication, and identity management in applications such as smartphones, border control, banking, and workplace access. A uni-biometric individual recognition system is an important module in most of the biometric systems. We propose automatic person recognition systems using both deep learning (DL) and machine learning (ML) techniques focusing on his voice. To achieve this goal, we propose two strategies. First, we customised YAMNet, a pretrained acoustic deep neural network, for individual speech recognition using a transfer learning technique. Second, we used transfer learning to shape YAMNet as a feature extractor for speech signals hybrided to a branch of ML algorithms as classifiers. The system was trained and tested with acoustic signals of speech in real environment. The classification results show that the proposed methods can perform an interesting rate of accuracy in nearly real time. The overall accuracy was 95.75% in frame level with the YAMNet-SVM model. The feature extractor-classifier established in this study provided a foundation for good behaviour biometric systems.

Keywords: Artificial intelligence; Deep learning; Machine learning; YAMNet; Biometric system; Individual recognition.

INTRODUCTION

In general, each human has his specific traits, such as facial features, iris patterns, fingerprints, speech characteristics and so on. The automatic biometric systems are based on the strong principle that those human traits are discriminant. Those traits can be used alone to construct uni-modal systems. However, two or more of those traits can be combined to feed multi-modal systems. Those systems are able to distinguish one person from another, with different levels of accuracy [1-4]. Compared to traditional authentication methods, biometric systems offer higher security, ease of use, and resistance to theft or loss. They were highly used during and after the COVID-19 pandemic [3].

Speaker recognition or biometric authentication of person based on their voice. It have an important role in various applications. The main goal is to extract unique vocal features that can distinguish one speaker from another. Traditional speaker recognition systems often relied on techniques such as Gaussian Mixture Models (GMMs), Mel-Frequency Cepstral Coefficients (MFCCs), and i-vector representations. While these methods achieved reasonable performance, they often struggled in real-world conditions with background noise, variable speech quality, limited training data, and implication of many compement such as vocabulary, acoustic model, language model, training algorithms and so on [5-7].

The artificial intelligence (AI) focuses on the principle of feeding automatic models with a significant quantity of data related to a general or specific topic in order to conduct automatic learning. Those obtained trained models are tested with unseen data to validate their ability to recognise new situations, according to the targeted tasks. The main categories of intelligent models are based on machine learning and deep learning. The main difference between them relies on the feature extraction step called the recognition patterns process of the available data. Recently, deep learning has brought significant improvements to the speech and speaker recognition fields. Deep neural networks can automatically learn pertinent features from raw audio or spectrograms, reducing the need for manual feature

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

extraction. Speaker recognition systems based on deep learning are becoming increasingly popular in both academic research and industrial applications [8].

YAMNet is a pretrained audio classification model incorporating the MobileNetV [8, 9] architecture that has been pre-trained on the Google AudioSet dataset of YouTube videos with over 1000 views [10]. To train or pretrain these huge networks with rich architecture, such as YAMNet, locally, it is necessary to have enormous amounts of data and powerful hardware material, which cannot be easily available in our case and in many real environments.

Recently, YAMNet is used for classifying a wide range of everyday sounds, making it valuable in applications like environmental sound detection, health monitoring, etc. In [11], the authors used acoustic signals recorded from bone conduction microphones as input and trained a model that combines transfer learning using YAMNet and the deep learning network Long Short-Term Memory (LSTM) to integrate them into the eating activity detection task. In [12], the YAMNet model was used as a feature extractor by outputting an intermediate layer embedding using transfer learning for COVID-19 cough classification.

This work focuses on two approaches for speaker identification based on the YAMNet model. First, we explore and implement the deep learning (DL) system baseline solution based on YAMNet, applying the transfer learning technique. Second, we consider a pipeline using the YAMNet as a feature selector and we hybrid 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) [13]. A hybridation of both of these techniques can improve accuracy and robustness. The main contributions in this work are as follows:

- Analyze the baseline performance of YAMNet across acoustique signal (Speech).

- Propose architectural improvements, such as the integration of four machine-learning algorithms (ML) as classifier to enhance recongnition accuracy of YAMNet.

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

This paper is composed of four sections. Section 2 presents the methodology and the background focusing on the DL and ML used in our proposed approaches and detail of the dataset. In Section 3, we describe and discuss the experimental results. Finally, we conclude this paper in Section 4.

METHODOLOGY

Before presenting the approach, we first introduce the overview of the key AI models that form the foundation of this work. This section also includes the essential details about the used dataset for feeding our models. Those details are presented in order to better follow the context and motivations behind the proposed approaches.

Models

YAMNet

Artificial neural networks are powerful models capable of learning complex nonlinear mappings between inputs and outputs. They are inspired from human biological neural network. The full YAMNet is a pretrained deep learning audio classification model, able to predict 521 different sounds providing from human, animals, things, etc. It is pre-trained on the Google/Youtube AudioSet dataset that contains over 632 different audio events [10]. YAMNet, which does not demand any feature extraction, learns the acoustic features from the input acoustic data with the specificities to be resampled into 16000 Hz in a single channel. The feature extraction or acoustic embedded layer converts the audio data into spectrograms, which are then used with MobileNet [9]. The full YAMNet inference process from sound file to prediction. Therefore, YAMNet is incorporating the MobileNet architecture. The YAMNet uses mel spectrograms as an input preprocessing form for audio signals. According to the architecture (seen in Table 1), the audio data needed as input for feeding YAMNet are resampled into 16000 Hz with single-channel audio, and the mel-scaled spectrograms are generated with 64 log-energies as a triangular filter bank.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Type	Filter shape	Input size
Input Layer	3×3×3	96×64×1
Conv$_1$	3×3×3	48×32×32
Conv$_2$dw	3×3×3 dw	48×32×32
Conv$_2$pw	1×1×32×64	48×32×64
Conv$_3$dw	3×3×64 dw	24×16×64
Conv$_3$pw	3×3×128	24×16×128
Conv$_4$dw	3×3×128 dw	24×16×128
Conv$_4$pw	1×1×128×128	24×16×128
Conv$_5$dw	3×3×128 dw	12×8×128
Conv$_5$pw	1×1×128×256	12×8×256
Conv$_6$dw	3×3×256 dw	12×8×256
Conv$_6$pw	1×1×256×256	12×8×256
Conv$_7$dw	3×3×256 dw	6×4×512
Conv$_7$pw	1×1×256×512	6×4×512
Conv$_8$dw-Conv$_{12}$dw	3×3×512	6×4×512
Conv$_{13}$dw	3×3×512	3×2×1024
Conv$_{13}$pw	1×1×512×1024	3×2×1024
Conv$_{14}$dw	3×3×1024 dw	3×2×1024
Conv$_{14}$pw	1×1×1024×1024	3×2×1024
Average Pooling (3×2)	1×1×1024	1×1×1024
Fully Connected (1024 ×512)	1024×512	1×1×512
Softmax (Classifier)	Confidence Scores	

Figure 1: The YAMNet architecture_ inspired from [9].

Support vector machines

Support Vector Machines (SVMs), proposed by Vapnik and Cortes in 1995, are supervised learning methods that need labelled data, designed primarily for binary classification tasks. The core principle behind SVMs is to identify the optimal hyperplane that maximizes the margin for a bi-classification task in a feature space, as illustrated in figure 1. SVMs rely on constructing a linear decision boundary in a transformed feature space using a mathematical tip called kernel functions. Initially developed for two-class problems, SVM was extended to multiclass classification

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

through strategies like one-vs-one or one-vs-rest approaches. SVM is considered as a cornerstone method in machine learning, with applications ranging from text classification and image recognition to bioinformatics [14].

Random Forest

The Random Forest (RF) regressor is a learning method combining many decision trees to improve predictive accuracy and robustness. RF constructs a collection of trees trained on different bootstrap samples (bagging) and randomly selects subsets of features at each split. For regression tasks, the RF aggregates the outputs of all trees by averaging their predictions [13].

Naïve Bayes

The naive Bayes algorithm (NB) is a simple and strong supervised ML classifier. In order to achieve a classification task, the algorithm computes the posterior probability for all classes and selects the one with the highest value, using likelihoods and prior probabilities of the features of the available data. The principal key assumption of the naive Bayes model is that the distributions of the input variables are conditionally independent [13].

K nearest neighbours

The K nearest neighbours algorithm (K-NN) first determines each data point's neighbourhood by identifying its K nearest neighbours or by locating every point inside a sphere with a specified radius. According to their Euclidean distance, all nearby points are connected and labelled [13].

Dataset

The VoxCeleb1 dataset is a large-scale audiovisual dataset primarily used for speaker recognition and verification. It includes over 100,000 English spoken sentences from 1,251 speakers, including both men and women. The total recording length is approximately 352 hours of speech. The voice recordings are automatically extracted from YouTube videos, including celebrity interviews. Although the data comes from audiovisual content, only the audio signals are retrained in this dataset. Each excerpt is annotated with a unique speaker ID, enabling tasks such as voice recognition and verification [16].

Proposed Approach

According to the proposed approaches, we first explore and implement the deep learning (DL) system baseline solution based on YAMNet with the transfer learning technique. Second, we consider a pipeline using the YAMNet as a feature selector and we hybridise each one of four machine learning (ML) classifiers, including Support Vector Machines (SVM), Random Forest (RF), K-nearest neighbours (K-NN) and Naïve Bayes (NBSupport Vector Machines (SVM), Random Forest (RF), K-nearest neighbours (K-NN), and Naïve Bayes (NB) are all examples. Figure 1 shows the global architecture of the proposed systems:

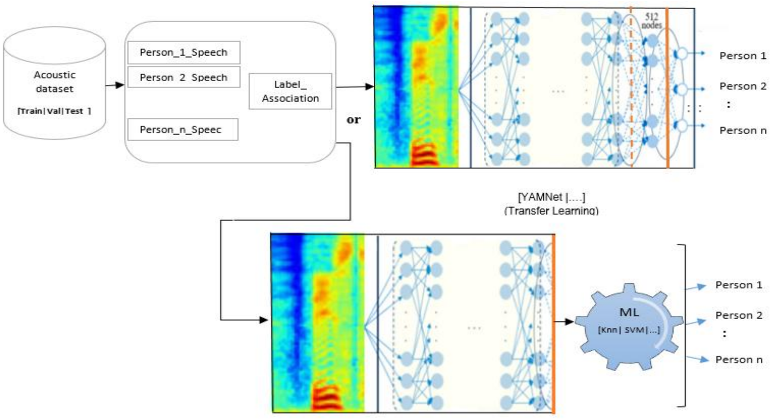

Figure 1: An illustration of the strategy applied to the proposed improved Sytem based YAMNet.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

RESULTS

This section shows the results of the proposed models in terms of accuracy and time processing. We will start with checking the baseline performance and then point out the improvements that we made with our hybridation based on ML algorithms. To determine a performance baseline, we used YAMNet with transfer learning according to the number of speakers in the acoustic dataset. We used the YAMNet deep learning approach for the first unimodal biometric system. This pre-trained model, based on the MobileNet architecture, allows efficient extraction of audio features. Speakers are then classified using the obtained descriptors. This method offers excellent 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 approximately 90.75%, demonstrating a good compromise between efficiency and low complexity. However, for more targeted tasks such as fine-grained speaker recognition, Table 2 summarizes the results obtained in terms of classification rates according to different epoch values:

#Epoch	Accuracy [%]	Time [s]
10	67.25	0.35
20	81.25	0.42
30	86.75	0.48
40	89.50	0.58
50	90.25	1.07
60	90.75	1.15
70	**90.75**	1.27

Table 2: Performance of Speaker recognition system based on YAMNet.

Table 3 shows that the choice of classification algorithm has a significant impact on system performance. The SVM (Support Vector Machine) algorithm achieves the best classification rate with 95.75%, indicating that it effectively exploits the features extracted by YAMNet, likely due to its ability to create optimal decision boundaries in a high-dimensional space. Random Forest ranks second with a rate of 86.75%, demonstrating good robustness, although slightly less efficient than SVM. This result is expected, as Random Forest can sometimes perform less well when the features are not highly discriminating or when the data is noisy. However, the K-NN algorithm (with cosine distance) achieves a rate of 81.75%, which remains acceptable but shows its limitations, particularly in terms of sensitivity to the local distribution of the data in the vector space generated by YAMNet. Finally, the NB algorithm gives the lowest classification rate (53.75%), which is explained by its simplifying assumptions (independence between characteristics), which are not realistic with the complex representations generated by YAMNet.

Model	Accuracy [%]	Time [s]
YAMNet$_{\_TransferLearning}$	90.75	1.27
YAMNet+SVM	**95.75**	2.05
YAMNet+RF	86.75	1.50
YAMNet+k-NN	81.75	2.00
YAMNet+NB	53.75	1.10

Table 3: Performance of Speaker recognition system based on YAMNet (DL) and Hybrided Systems (DL +ML).

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

CONCLUSION

The automatic biometric systems are able to distinguish one person from another, with good level of accuracy. We propose automatic person recognition systems using both Deep Learning (DL) and hybridation of DL and Machine Learning (ML) techniques focusing on his voice. In order to achieve this goal, we propose two strategies. First, we adapted YAMNet, the acoustic deep neural network, for individual speech recongnition using transfer-learning technique. Second, we used transfer learning on YAMNet as a feature extractor for speech files hybrided to a branch of ML algorithms as classifiers.

In this study, the acoustic recongition performance of individual has been evaluated, providing valuable insights for practical applications and future researchs. The evaluation has been performed using two senarios. First, we have propose to use the deep neural network YAMNet by appling a transfer learning. Second, we use the specific layers of YAMNet for extracting adequate features in the frequency domain and then we apply many machine-learning algorithms. The results show good recognition results are achieved of 95.75% by the YAMNet_SVM hybrid model.

As future work, we plan to explore other deeper transformer architectures and other machine learning algorithms, particularly for individual recognition in noisy environments. We focus on achieving more improvement in terms of accuracy and time processing. Another point that we consider is to increase the data size for more extended comparisons.

REFRENCES

[1] 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).

[2] 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.

[3] 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.

[4] 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.

[5] Rabiner, L. R., & Schafer, R. W. (2011). Theory and applications of digital speech processing. Pearson Education.

[6] 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.

[7] 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.

[8] K. K. Mohammed, E. I. Abd El-Latif, N. Emad El-Sayad, A. Darwish, A. Ella Hassanien. Radio frequency fingerprint-based drone identification and classification using Mel spectrograms and pre-trained YAMNet neural. Internet of Things, 23:100879, 2023. Elsevier.

[9] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

[10] Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., ... & Ritter, M. (2017, March). Audio set: An ontology and human-labeled dataset for audio events. In 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 776-780). IEEE.

[11] Chen, W., Kamachi, H., Yokokubo, A., & Lopez, G. (2022, January). Bone Conduction Eating Activity Detection based on YAMNet Transfer Learning and LSTM Networks. In BIOSIGNALS (pp. 74-84).

[12] Elizalde, B. and Tompkins, D. (2021). Covid-19 detection using recorded coughs in the 2021 dicova challenge. arXiv preprint arXiv:2105.10619.

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

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

[14] Cortes, C. and Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3):273‐297.

[15] Mahum, R., Irtaza, A., Javed, A., Mahmoud, H. A., & Hassan, H. (2024). DeepDet: YAMNet with BottleNeck Attention Module (BAM) for TTS synthesis detection. EURASIP Journal on Audio, Speech, and Music Processing, 2024(1), 1

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

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.