Deep Speech Recognition System Based on AutoEncoder-GAN for Biometric Access Control

Oussama Mounnan[1]
LABSI Laboratory, Faculty of Sciences
Ibn Zohr University Agadir, Morocco
LIASD Laboratory, Saint-Denis, France

Otman Manad[2]
Umanis S.A Research & Innovation,
7 Rue Paul Vaillant Couturier
92300 Levallois-Perret, France

Abdelkrim El Mouatasim[3]
Department of Mathematics and Management
Faculty of Polydisciplinary Ouarzazate (FPO)
Ibn Zohr University, Ouarzazate, Morocco

Larbi Boubchir[4]
LIASD Laboratory
Department of Computer Science
Paris 8 University, 93526 Saint-denis, France

Boubaker Daachi[5]
LIASD Laboratory
Department of Computer Science
Paris 8 University, 93526 Saint-denis, France

***Abstract*—Speech recognition-based biometric access control systems are promising solutions that have resolved many issues related to security and convenience. Speech recognition, as a biometric modality, offers unique advantages such as user-friendliness and non-intrusiveness, etc. However, developing robust and accurate speaker identification and authentication systems pose challenges due to variations in speech patterns and environmental factors. Integrating deep learning techniques, especially AutoEncoder and Generative Adversarial Network models, has shown promising results in addressing these challenges. This article presents a novel approach based on the combination of two deep learning models, namely, AE and GAN for speech recognition-based biometric access control. In the model architecture, the AutoEncoder takes the MFCC coefficients as input, and the encoder converts the latter to the latent space, whereas the decoder reconstructs the data. Then, speech features extracted from the latent space are used in the GAN generator to generate additional speech data. The discriminator network has a dual role, serving as both a feature extractor and a classifier. The first extracts relevant features from generated samples, while the latter distinguishes between generated and authentic samples that come from AutoEncoder. This strategy outperforms DNN and LSTM models on VoxCeleb 2, LibriSpeech, and Aishell-1 datasets. The models are trained to minimize Mean Squared Error (MSE) for both the generator and discriminator, aiming at achieving highly realistic datasets and a robust, interpretable model. This approach addresses challenges in feature extraction, data augmentation, realistic biometric samples generation, data variability handling, and data generalization enhancement, providing therefore, a comprehensive solution.**

Keywords*—*Speaker identification; speech recognition; biometric access control; authentication; verification

I. Introduction

Speech recognition systems [1] have become increasingly important in various domains, including biometric access control, where the identification and authentication of individuals based on their unique voice characteristics are crucial. These systems aim, securely, to use biological or behavioral characteristics to authenticate and authorize individuals for access to a physical location, a device, or a system. It relies on unique and measurable traits that are specific to an individual, making it difficult to forge or replicate. These characteristics can include physiological characteristics such as fingerprints, face features, iris patterns, and voiceprints, as well as behavioral characteristics such as typing patterns, gait, and signature dynamics as shown in Fig. 1. The main function of this paradigm is to collect biometric data, convert it into a digital template, and then compare this template to templates stored in a database. If the comparison results in a match, the individual is given access. Otherwise, access is blocked. Fig. 2 presents a system architecture based on speech recognition.

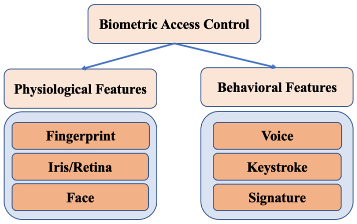

Fig. 1. Biometric categories.

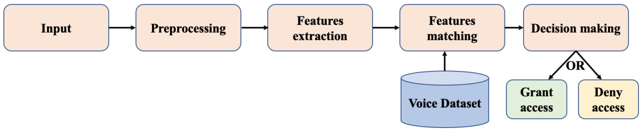

Fig. 2. Biometric access control architecture based on speech recognition.

Biometric access control systems based on speech recognition offer numerous advantages [2], such as universality, non-intrusiveness, high authentication security, and convenience i.e., bypassing the use of memorized passwords or access cards, audit trail (keep track and accountability), and faster processing than the traditional one. This concept is a powerful and convenient way to improve security and access management in a variety of areas, such as physical facilities, digital systems, and digital transactions. . . etc. However, to achieve reliable and robust performance, it is essential to develop accurate speaker identification and authentication mechanisms.

Deep learning models are characterized by their ability to learn sophisticated patterns and representations from data, providing a strong foundation for tackling the complexities of speech recognition [3]. Hence, providing powerful tools for improving the accuracy and effectiveness of its tasks. In this context, the use of AutoEncoder models for feature extraction [4] has shown promising results in identifying and authenticating speakers. The AutoEncoder model, a type of neural network architecture, has gained significant attention in recent years due to its capability to learn meaningful and compact representations of input data. AutoEncoders can be leveraged to extract discriminative features from raw speech signals. By training the AutoEncoder model on a large dataset of labeled speech samples, it can learn to encode the essential characteristics of a speaker's voice into a lower-dimensional representation, which facilitates efficient and accurate speaker identification and authentication processes. However, this model faces several challenges among them:

- Lack of Realism in Generated Samples: Because AutoEncoders concentrate on recreating the input data, they may produce generated samples that are excessively similar to the training data and are devoid of variation.
- Noisy or incomplete reconstructions: AutoEncoders may have trouble accurately reconstructing the input speech signals, mainly when there is noise or fluctuation.
- Limited Generalization of novel data: Because AutoEncoders tend to concentrate on recreating well-known patterns from the training set, they may have trouble in generalizing novel or unseen data.
- Incapability to Distinguish: AutoEncoders are typically unsupervised models concentrating on feature learning and reconstruction. The capacity to discriminate is essential for precise authentication in a biometric access control context.
- Limited data augmentation: AutoEncoders can be used for limited data augmentation by reconstructing and producing synthetic samples. The produced samples, however, may not represent the whole range of variability included in the training data.
- Adversarial Attacks: AutoEncoders can be vulnerable to adversarial attacks [5], which include making tiny and purposeful changes to input data in order to trick the model. In the case of speech recognition systems, this might include discreetly changing a voice recording to deceive the system into providing access to an unauthorized user.
- Inadequate Temporal Information: Traditional AutoEncoders struggle with sequential data, which is an issue in speech recognition, where the order of the input (i.e., the sequence of sounds or words) is important. Recurrent or convolutional AutoEncoders, for example, can alleviate this, although they are more sophisticated and computationally intensive.

Generative Adversarial Network (GAN) is a promising paradigm that consists of two main components: a generator and a discriminator. The generator produces synthetic data and the discriminator tries to differentiate between real and synthetic data. In the context of biometric access control using speech recognition [6], GANs can be applied to generate synthetic speech data to augment the training dataset [7], which can help address data scarcity issues, increase the diversity of the training data, and improve the robustness and generalization of the speech recognition system. Synthetic data generated by the GAN can be combined with real training data to create a more comprehensive and representative dataset for training speaker identification or authentication models, mitigating most AE model issues. It's important to note that GAN training can be challenging and may suffer from issues [8] such as training instability requiring careful tuning of hyperparameters and balancing the training dynamics between the generator and discriminator, lack of control over generated data, GANs typically generate data based on random noise input, resulting in limited control over specific characteristics of the generated speech samples, lack of feature extraction: GANs primarily focus on generating data and may not explicitly learn meaningful features from the input speech samples, and data augmentation: GANs are commonly used for data augmentation by generating synthetic samples. However, without the guidance of meaningful features, the generated samples may not effectively capture the desired variations and characteristics of real speech data. The combination of AutoEncoder (AE) and Generative Adversarial Networks (GAN) models, for biometric access control based on speech including speaker identification and authentication, resolves many problems and drawbacks related to feature extraction, data augmentation, and generating realistic biometric samples. To this end, the main goal is to contribute to the advancement of this research by leveraging the capabilities of deep learning through a novel approach that combines both models in a manner complementary to each other, providing control, stability, better feature learning, and enhanced data augmentation capabilities, leading to improved speech recognition performance.

The key contributions of this project can be outlined as follows:

- Introducing a novel approach rooted in deep learning models, specifically AutoEncoder and Generative Adversarial Network in biometric access control through speech recognition context. This integrated model enhances system performance, accuracy, robustness, and efficiency.
- Employing the AutoEncoder (AE) model as an unsupervised method for extracting meaningful and discriminative features, reducing dimensionality, and addressing storage and computational challenges associated with raw audio data analysis. Additionally, features are extracted from the latent space and utilized in the Generative Adversarial Network (GAN) to augment the training dataset, enhancing model generalization, mitigating overfitting, and alleviating data scarcity issues, resulting in the creation of high-quality, realistic biometric samples.
- Proposing a GAN model where the generator network produces synthetic speech data resembling that from the latent space representation, thereby expanding

the training dataset. This approach improves model generalization, reduces overfitting, and addresses data scarcity, leading to the generation of high-quality biometric samples. The discriminator in this proposal serves two roles: feature extraction and classification. The former extracts features from generated samples, capturing more informative and efficient features, while the latter distinguishes between generated samples from both models, enhancing overall system performance.

- Application of this approach to diverse datasets, including VoxCeleb 2, Aishell-1, and LibriSpeech, has yielded positive results when compared to outcomes from Deep Neural Network (DNN) and Long Short-Term Memory (LSTM) models.

The remainder of this article is organized as follows: Section II provides related works on speech recognition systems for biometric access control. Section III presents the proposed solution. Section IV presents the results and analysis of the experiments conducted, highlighting the performance gains achieved through the AutoEncoder-GAN-based approach. Section V presents a discussion. Finally, Section VI concludes the article with a summary of the findings and discusses potential directions for future research in this field.

II. Related Work

In the literature, there is a lot of research related to biometric access control based on speech recognition topics, including speaker identification and authentication, and speaker verification. This section presents an overview of some works and propositions published recently that achieved significant results.

Najim Dehak et al [9] proposed two speaker verification systems models, In which the first one is based on SVM, by using the cosine similarity, and the second one utilises directly the cosine similarity in the final phase which decides the final score. The experiments are done through three different methods in the variability space, which are within-class covariance normalization, linear discriminate analysis, and nuisance attribute projection. Their study conducted on the combination of LDA with WCCN has achieved good results compared to the other ones. The test was carried out using the NIST 2008 Speaker Recognition Evaluation dataset.

Yen Lei et al [10] presented a new approach based on deep learning speaker recognition using a phonetically that aims at improving speaker recognition performance by using an i-vector model [11] to represent the speech signal (extract the main features) and DNN model is used to replace the UDM-GMM [12] paradigm in order to train the model. The experiments proved that this approach has significantly improved the i-vector speaker recognition system.

Another research done by [13] has proposed d-vector instead of i-vector that aims at extracting hidden layers of a DNN as features. D-vector represents the averaged activations from the last hidden layer of DNN. Experiments of this approach have proved its efficiency in a small-footprint text-dependent speaker task. Generally, this scheme underperforms the predecessor based on i-vector-DNN.

Another research made by [14] has proposed a multi-task deep learning scheme based on the j-vector method that consists of extracting features from multitask DNN using probabilistic linear discriminant analysis (PLDA). This scheme has achieved good results than the predecessor models (i-vector, d-vector).

The Authors in [15] have proposed a new scheme based on deep neural network DNN to extract speech features called as x-vector. This latter represents the fixed-dimensional embeddings of variable-length traits. Furthermore, this research tackled also data augmentation by adding the noise and the reverb to the existing dataset to improve the efficiency of the model in the text-independent speaker tasks. Effectively, this approach has achieved better findings than the ones based on the i-vector and d-vector. Another research conducted by [16] has proposed a new end-to-end architecture based on neural networks, especially DNN and LSTM to speaker verification in the text-dependent context that aims at mapping the utterances to a score and joining them to optimize the representation of the speaker. In the same area, the authors of [17] have proposed another approach based on the end-to-end attention model. They use the CNN model to extract the noise-robust frame-level features that will become utterance-level speaker vectors using the attention model. This approach proves its effectiveness on Windows 10 “Hey Cortana”.

Another research carried out by [18] in the context of text independence has presented a new end-to-end approach based on the deep learning model to optimize the triple loss function using Residual Net block and measuring the similarity by Euclidean distance within trials. The findings show that this approach outperforms that based on conventional i-vector schemes, namely on short utterances.

In [19], the authors have proposed a new generalized End-to-end model based on LSTM. The training process has relied on the large number of utterances forming a batch. This scheme aims at optimizing the loss function through the training process in an efficient manner. The experiments show that this platform has achieved good results. N. Le et al [20] have proposed a new approach based on deep learning model, namely CNN. The main objective of this proposition is to optimize the deep speaker embedding through intra-class loss distance variance regularization compactness. The findings have proved that this approach accelerates the convergence of the training model, which enhances the model’s performance.

Another research carried out by the authors in [21] has presented an end-to-end optimized scheme based on deep convolutional features extractor combined with self-attentive and large-margin loss functions in the text-independent tasks context. They use a modular neural network instead probabilistic linear discriminant analysis (PLDA) classifier. This work made use of the experiments on VoxCeleb and NIST-SRE 2016 and has achieved an enhancement model than the others based on i-vectors.

The authors in [22] have suggested a novel approach for learning speaker embeddings based on a simulated model of GAN, especially the discriminator. This architecture aims to maximize mutual information, improving the model performance on the VoxCeleb corpus. Experiments show that this model outperforms the model based on i-vector and that based on triples loss systems.

Many works are proposed to optimize the performance of speech recognition tasks and provide a robust system using deep learning model. Each research has focused on one aspect or more, such as data augmentation, features extraction, denoising and de-reverberation. The proposed solution has designed a new architecture based on deep learning models, namely AutoEncoder and Generative Adversarial Network in a complementary manner to improve the model performance by minimizing the loss function. The MFCC is used to extract features and the model AE to capture the meaningful speech representation and GAN is used to generate speech data from the latent space of the model AE.

III. Proposed Scheme

The proposed scheme is based on two models which are AE and GAN models as depicted in Fig. 3. At first, the speech inputs are collected, and their Mel-Frequency Cepstral Coefficient (MFCC) characteristics are extracted and used for training and tuning the model. Generally, The AE model comprises three components: Encoder, latent space, and Decoder. The encoder captures the main representation of the meaningful speaker speech features extracted from MFCC and produces the latent space, the latter will be used to reconstruct the input data. In this model, the Latent space will be extracted and used as input to the generator of the GAN model to generate more real data from it. The generated samples will be then used as input to the discriminator. This latter plays two roles, namely a features extractor and a classifier. At first, the discriminator extracts features from the generated samples and then feeds to the classification between that extracted and that comes from the AutoEncoder i.e. the decoder, to make a decision. This section presents more details of this model.

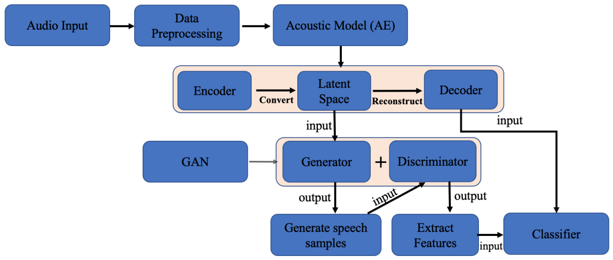

Fig. 3. AutoEncoder-Generative adversarial network model architecture.

A. Data Preprocessing

Generally, the preprocessing process [23] is crucial for preparing the data. This phase involves the capture and splitting of data into segments, feature extraction, noise removal, features normalization, and data loading, etc. Among the main steps that represent the backbone of the model namely in the context of the biometric access control based on speech, is feature extraction. To this end, the proposed architecture involves the adoption of the Mel-frequency Cepstral coefficients (MFCC).

1) Mel-Frequency Cepstral Coefficients: MFCC [24] is a technique that consists of extracting features from the signal. In the speech processing context, this method is widely used to capture the spectral features of sound well-suitable for various machine learning and deep learning tasks including speech recognition and speech analysis. Simply this technique is an amount of coefficients that represent the shape of the speech power spectrum signal. Fig. 4 represents the components of the MFCC. To calculate the coefficients of MFCC, some steps are crucial as depicted in the figure. After capturing the speech signal, the first step is breaking the signal into frames (windowing process) and then applying the Fast Fourier Transform (FFT) to determine the power spectrum of each frame. Following that mel-scale filter bank processing is performed on the power spectrum by the formula 1:

$$mel(f) = 2595log_{10}(1 + \frac{f}{700}) \quad (1)$$

Where mel(f) represents the frequency on mels and f represents the frequency on Hz. The power spectrum is converted then by log domain and the Discrete Cosine Transform (DCT) is applied to get the coefficients of MFCC through the Eq. 2:

$$\hat{C}_n = \sum_{n=1}^{k}(log\hat{S}_k)cos[n(k-\frac{1}{2})\frac{\pi}{k}] \quad (2)$$

Where $k, \hat{S}_k, and\hat{C}_n$ represent, respectively the mel cepstrom coefficients numbers, the filter bank output and the MFCC coefficients.

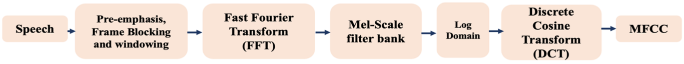
Speech
Pre-emphasis, Frame Blocking and windowing
Fast Fourier Transform (FFT)
Mel-Scale filter bank
Log Domain
Discrete Cosine Transform (DCT)
MFCC

Fig. 4. MFCC architecture.

B. AutoEncoder model

The AutoEncoder (AE) model is a sort of neural network architecture used for feature learning, dimensionality reduction, and data reconstruction. It is especially effective for extracting relevant representations from biometric data and may be used in a wide range of biometric modalities, in various applications, namely Feature learning, data denoising, data compression, Anomaly detection, Privacy preservation, biometric template protection...etc. An AutoEncoder's primary principle is to learn a compact and efficient representation of incoming data. It is made up of an encoder and a decoder as shown in Fig. 5. The encoder takes raw data as input and converts it to a lower-dimensional latent space representation. The fundamental traits and qualities of the data are captured by this latent space representation, and the decoder uses this later to attempt to recreate the original input data. The objective is to maintain the information required for reconstruction in the latent space. AE is an unsupervised model that aims at minimizing the loss function between the input data and the reconstructed data, capturing the most relevant representations.

The proposed solution incorporates the use of AutoEncoder to capture the relevant representation of the inputs from MFCC coefficients, optimizing the speech processing system. The main objective of MFCC is extracting features and converting the input signal into coefficients that are retained as features which represent the main relevant features. The AutoEncoder takes these coefficients as input and converts them into latent space, reducing therefore, the dimensionality of the representation represented by the coefficients, and extracting the main relevant representation. The other network i.e. decoder network reconstructs the representation from that reduced (latent space). The main goal of this proposition is to get the most salient and compact representation from MFCC coefficients in a lower-dimensional space by training the AutoEncoder model.

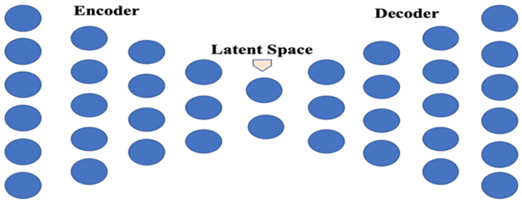

Fig. 5. AutoEncoder architecture.

C. GAN model

Generative Adversarial Network (GAN) is a generative model that is distinguished by two distinct networks, each with its unique set of attributes called Generator and Discriminator. The first seeks to produce realistic data from a specific class, while the second is used to determine whether the generated data is realistic or phony, as shown in Fig. 6: GAN is a deep-learning class used especially to produce synthetic data from the raw data input. In the scope of biometric access control, the generator takes a random noise as input and attempts to produce biometric data samples that mimic actual biometric data, while the objective of the Discriminator is to distinguish between the generated samples and the real ones, generally a binary classifier. The training procedure comprises a competition between the generator and the discriminator. As training advances, the generator improves at creating more realistic data, while the discriminator improves at differentiating between actual and phony data. This repeated procedure should result in high-quality synthetic data that is difficult to differentiate from genuine data. GANs may be used for a variety of reasons in the context of biometric access control, including data augmentation, Privacy-Preserving Research, Training Data Generation, Data Imputation, and Adversarial Attacks and Defense.

To this end, the proposed scheme extracts the latent space from the AutoEncoder model and uses it as input in the Generative Adversarial Network (GAN) model namely the Generator. This latter Generates more speech data from those reduced features, producing then data simulated to that of input. Whereas the discriminator in this architecture plays two roles, namely a features extractor and a classifier. At first, the discriminator takes the generated samples as input, extracts relevant features and then distinguishes them from that produced and trained by the AE model, especially the decoder.

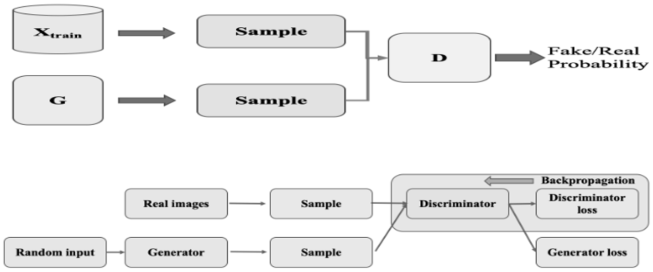

Fig. 6. GAN architecture.

IV. Experiments and Results Analysis

In this section, the experiments carried out by the laboratory team are presented, describing therefore, the datasets, the metrics and the implementation details of the proposed model, and finally the analysis of the result.

A. Datasets

In this model, three different datasets have been used, which are VoxCeleb 2, Aishell-1 And LibriSpeech.

VoxCeleb: It is an open-source dataset that is widely used in the experiments of speech processing tasks [25]. It contains videos interviews uploaded to YouTube. There are two types of VoxCelb datasets, VoxCeleb 1 and VoxCeleb 2. The first one has over 100,000 utterances For celebrities, whereas the VoxCeleb 2 has over a million utterances. In the proposed solution, the experiments have occurred on the VoxCeleb 2.

Aishell-1: This dataset is used also in the speech preprocessing tasks[26]. It is an open-source and freely accessible speech dataset that contains Mandarin speech captured with a high-fidelity microphone (44.1 kHz, 16-bit). The Aishell-1 dataset was created by downsampling the audio collected by the high-fidelity microphone to 16 kHz. A set of 400 speakers from various accent areas in China took part in the record capture.

LibriSpeech: is an open-source corpus, available in [27]. It contains 1000 hours of speech sampled at 16KHz and is generated from audiobooks in the LibriVox project. This dataset is used mainly in speech preprocessing including speech recognition and speaker identification. Table I represents the specification of the used datasets:

TABLE I. Datasets Specification

	Speakers	Utterances	Hours
VoxCeleb 2	6112	1 128 246	2442
Aishell 1	400	141 925	Over 170
LibriSpeech	2087	252 702	1000

B. Evaluation Metrics

Generally, a metric is a method used to evaluate a system's performance on a specific task. The main metric objective is measuring the quality of classifications or predictions carried out by a system or model. A loss or error function [28] is a function that determines how much the output or predicted value departs from reality or actual value aiming at optimizing the model (either maximizing or minimizing issues). The Mean Squared Error (MSE) [29] is a loss function that measures the error between the observed and predicted values. The average of errors squared is calculated by this Eq. 3.

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n} \tag{3}$$

Where y_i represents the observed value, $\hat{y}_i$ represents the predicted value, and n is the observations number In this study, the MSE metric is used in the AutoEncoder model to evaluate its performance, and the Binary Cross Entropy (BCE) metric is also used in the GAN model that represents the difference between the predicted probability distribution and the reel one. On one hand, BCE is used to solve the binary classification issues, evaluating then, the model's performance. On the other hand, is used to quantify the training loss, minimizing therefore, the loss function of the model during training.

C. Implementation Configuration and Results Analysis

In the proposed scheme, the PyToch library, written in Python programming language, has been used for training networks based on deep learning models. This model has adopted Graphics Processing Unit (GPU), due to its efficiency in Neural Network processing. After capturing the MFCC coefficients from the speech, 13 coefficients, The latter are then fed to the AE model namely, the encoder that converts the inputs to latent space, reducing therefore the dimensionality and capturing essential speech features. These high-level features serve to reconstruct the speech data from the bottleneck layer and aim at generating outputs that closely resemble the original data input. The model training aims at minimizing the reconstruction loss between the original and reconstructed speech. In the training of the AutoEncoder, the chosen specifications include 8 dimensions as the latent dimension, 64 as the batch size, and 0.001 as the learning rate. Within the first part of the proposed architecture, the AE model is implemented with input dimensions set to 8. The encoder network consists of 128 units or neurons in the hidden layer, employing the Rectified Linear Units (ReLU) as an activation function. The use of ReLU introduces non-linearity, facilitating the model in learning complex relationships within the data. The learning rate and the network size are identified using different settings based on the try-and-error approach to choose the best configuration in terms of performance.

At the beginning of the training process, the weights are initialized at random and then gradually updated. To solve the model overfitting challenges, different methods are used such as the regularization of the parameters to promote lower values of weight, and adding dropout layers within the encoder and the decoder, furthermore, the loss function regularization has been adopted to promote certain desired behaviors in the latent space. The data mapping process is carried out from the 128-dimensional hidden representation to the latent space representation. The Adam optimizer has been deployed. The loss function is selected as the Mean Squared Error (MSE) as mentioned before.

In this proposition, the latent space features are extracted representing the high-level speech representation to feed it into the GAN model, namely the generator network. This latter takes the high-level representations (more relevant speech features) as input to generate more speech data in a manner that resembles real speech. The architecture of the generator is composed of three fully connected linear layers with ReLU activation functions between them. The Tanh activation function has been applied in the final layer to ensure that the generated values are bounded within the range [-1,1]. The other GAN network, i.e., the discriminator, plays two roles in this architecture, a features extractor and a classifier. At first, the discriminator takes the generated samples from the generator, tries to extract the relevant representations and then feeds them to the classifier to distinguish them from those that come from the decoder of the AE model. The structure of the discriminator is similar to that of the generator. It consists of three fully connected layers with ReLU activation functions. In the final layer, the sigmoid activation function has been applied, which produces values within the range [0,1] where 1 identifies the real data, and 0 identifies the fake ones. Both the generator and the discriminator are adversarial trained. i.e. competing against each other. This process helps us to refine the ability of the generator to generate more high-level quality speech data, and therefore, achieve a robust system based on the combination of two promising deep learning models, AutoEncoder and Generative Adversarial Network, especially in the speech recognition tasks. Fig. 7 represents AE-GAN model training process using three different datasets, with the loss versus training epochs to illustrate how well the model learns. The experiments incorporate different utterances from three different datasets, including VoxCeleb 2, LibriSpeech, and Aishell-1. These datasets are divided into three parts for each dataset, 80% for training, 10% for validation, and 10% for test.

Experimentation involved assessing the proposed deep AE-GAN model by utilizing the state-of-the-art models, namely the Deep Neural Network (DNN) model and Long Short Term Memory (LSTM) model, using the datasets mentioned above to describe the experiment findings. Table II lists the overall loss function of the models that are used in the test process during various research phases. As shown in the table, the AE-GAN model has a high score in training and validation in three different datasets, which are LibriSpeech, VoxCeleb 2, and Aishell 1, it has achieved respectively in training loss, 0.0574, 0.0876, and 0.0886, and in validation loss 0.0581, 0.0888, and 0.0889. Compared to the results of DNN and LSTM models, they have gotten in the training phase values ranging from 0.07 and 0.168, while in the validation phase, huge values ranging between 0.30 and 0.48, proving generally the overfitting of the models. The proposed scheme has proved its efficiency and outperformed the performance of DNN and LSTM models in three different datasets. Fig. 8 depicts the results of the experiments carried out over the datasets using the DNN and LSTM models.

V. Discussion

Deep Neural Networks (DNNs) and Long Short-Term Memory networks (LSTMs) are reference models in speech recognition-based biometric access control context, and have been widely used in many studies. DNNs have demonstrated their performance in learning hierarchical representations from raw audio data. Their ability to handle complex features with

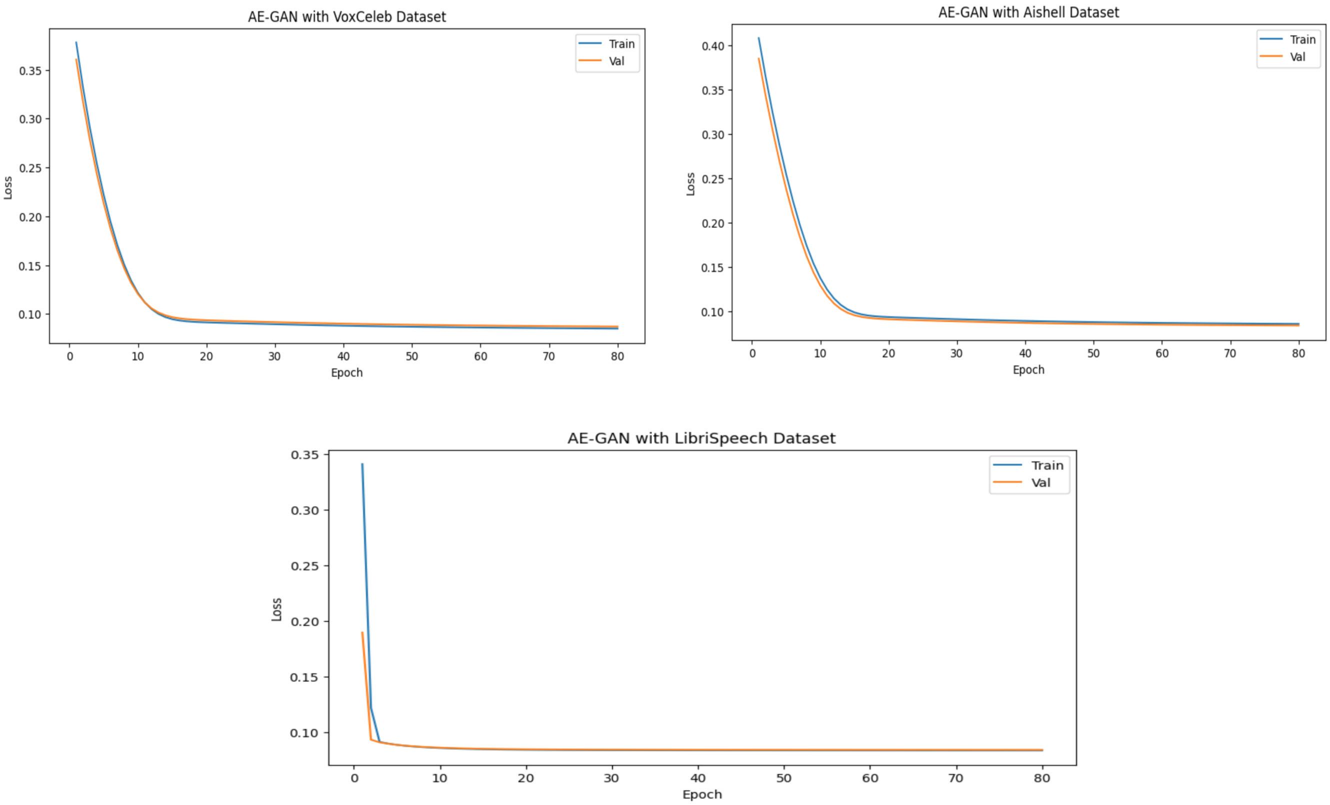

Fig. 7. AE-GAN model training loss curve vs epoch.

TABLE II. AVERAGE LOSS PER EPOCH FOR TRAINING AND VALIDATION

Dataset	Model	Loss function	
		Train	Validation
	DNN	0.079	0.334
LibriSpeech	LSTM	0.095	0.307
	AE-GAN	**0.0574**	**0.0581**
	DNN	0.152	0.298
VoxCeleb 2	LSTM	0.168	0.287
	AE-GAN	**0.0876**	**0.0888**
	DNN	0.084	0.389
Aishell 1	LSTM	0.079	0.486
	AE-GAN	**0.0886**	**0.0889**

deep increased network has contributed to their standing in the field. However, this model struggles with capturing long-range dependencies in sequential data, which is crucial for speech recognition tasks. Simultaneously, LSTMs are recognized for their effectiveness in modeling temporal dependencies within sequential data, making them well-suited for capturing long-term patterns in speech sequences. They have addressed the vanishing gradient issues that are inherent in traditional Recurrent Neural Networks (RNNs), making them more adept at learning from sequential data. But to capture complex patterns effectively, LSTMs require more data.

The AE-GAN's ability to leverage the latent space features extracted by the AutoEncoder to enhance the generative capabilities of the GAN is a potential advantage. In scenarios with limited labeled data, the AE-GAN's capacity for generating high-quality and realistic speech samples may prove advantageous. Additionally, its ability to address overfitting challenges through regularization techniques and dropout layers may contribute to superior performance in diverse speech recognition tasks. Although the proposed model offers several advantages, it may face challenges in scenarios where there is insufficient diversity in the training data, potentially leading to biased representations. If the dataset lacks sufficient variation in terms of speakers, accents, or speech characteristics, the model may struggle to generalize well to a broader range of real-world scenarios. Augmenting the dataset with more diverse samples could enhance the model's robustness. Additionally, the model's performance may be sensitive to hyperparameter settings, necessitating careful tuning. Implementing automated hyperparameter tuning methods or conducting a thorough sensitivity analysis may help identify robust configurations more efficiently. The computational complexity of the model, especially in training large-scale datasets, could pose limitations in terms of time and resource requirements. The computational costs related to the training deep learning models, including the proposed AE-GAN model, are a significant consideration.

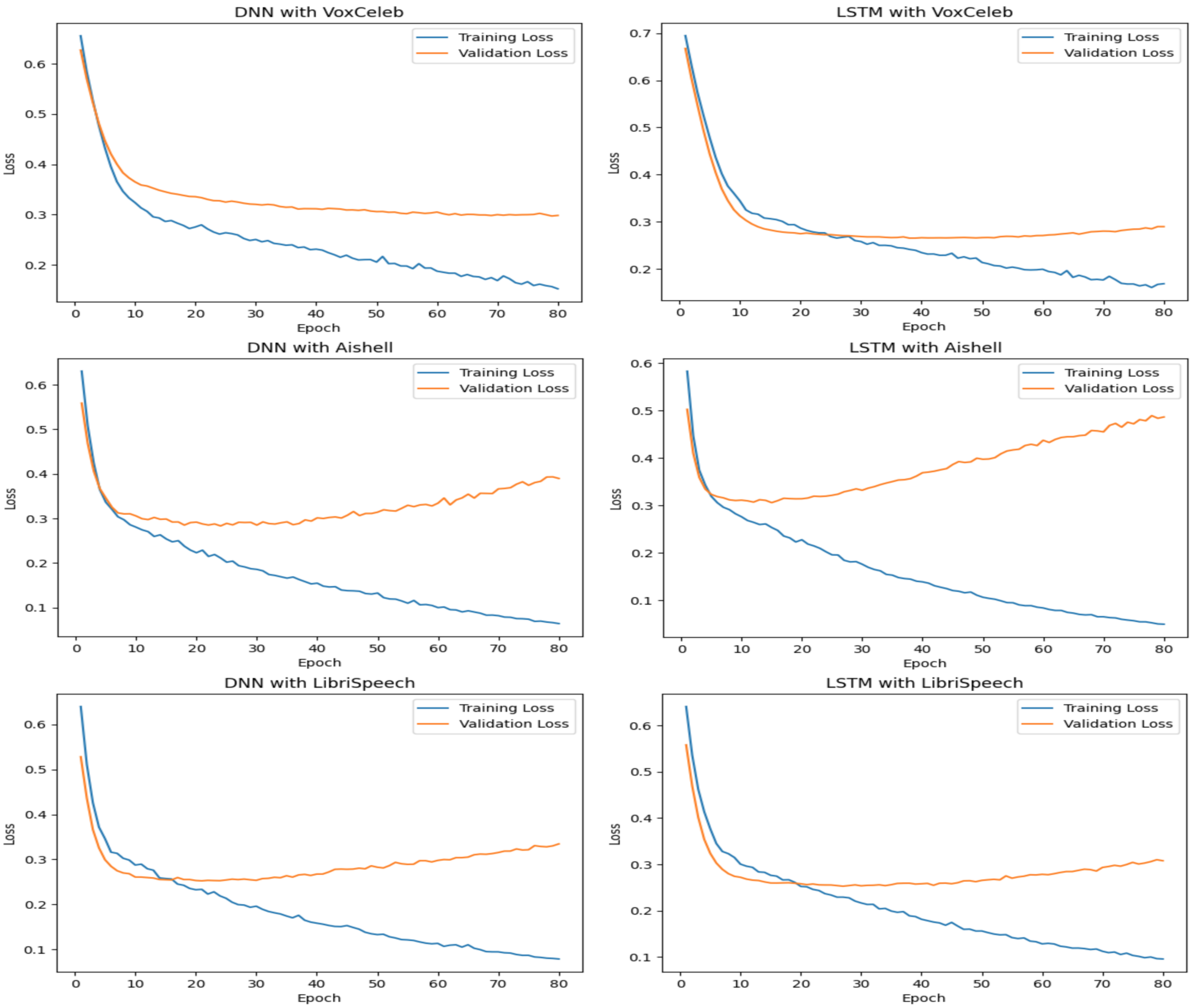

Fig. 8. The results of DNN and LSTM models with VoxCeleb, aishell, and libriSpeech datasets.

There are several solutions that contribute to mitigating these costs among them, efficient GPU utilization can be essential, optimizing the model architecture and exploring parallelization techniques can contribute to faster training times. Additionally, the use of transfer learning from pre-trained models can alleviate the need for extensive training on large datasets. Another potential solution is the exploration of model quantization techniques, reducing the precision of model weights to accelerate inference. Leveraging distributed training across multiple GPUs or utilizing cloud-based computing resources can further expedite the training process. Implementing early stopping and model checkpointing strategies can optimize training efficiency by preventing unnecessary iterations. Generally, a nuanced understanding of the AE-GAN model's strengths, a transparent acknowledgement of study limitations, and proactive strategies to address computational costs collectively contribute to a robust evaluation framework for advancing the field of speech recognition.

VI. Conclusion and Future Work

This paper has proposed a new approach based on speech recognition for speaker identification and authentication that is considered as the main and crucial task in the speech-based biometric access control scenario. The model has proved its efficiency and robustness based on the combination of AE and GAN models. The proposed model provides an optimized platform integrating the features learning and tackling the data augmentation and generalization issues, especially the speech dataset, and data imputation such as reconstructing degraded audio or denoise and tuning the hyperparameters of the models. This approach has been implemented on three different datasets: VoxCeleb2, LibriSpeech, and Aishell-1, and has achieved good results in terms of performance, compared to AE and GAN models. However, the proposed scheme is expensive in terms of time-consuming, especially in the training phase where there are two models AE and GAN. In future endeavors, the focus will be on this aspect to optimize the proposed scheme.

References

[1] S. Li, J. You, and X. Zhang, "Overview and Analysis of Speech Recognition," in 2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA), Aug. 2022, pp. 391–395. doi: 10.1109/AEECA55500.2022.9919050.

[2] V. S. S. S. Hari, A. K. Annavarapu, V. Shesamsetti, and S. Nalla, "Comprehensive Research on Speaker Recognition and its Challenges," in 2023 3rd International Conference on Smart Data Intelligence (ICSMDI), Trichy, India: IEEE, Mar. 2023, pp. 149–152. doi: 10.1109/ICSMDI57622.2023.00034.

[3] K. B. Bhangale and M. Kothandaraman, "Survey of Deep Learning Paradigms for Speech Processing," Wireless Pers Commun, vol. 125, no. 2, pp. 1913–1949, Jul. 2022, doi: 10.1007/s11277-022-09640-y.

[4] O. İrsoy and E. Alpaydın, "Unsupervised feature extraction with AutoEncoder trees," Neurocomputing, vol. 258, pp. 63–73, Oct. 2017, doi: 10.1016/j.neucom.2017.02.075.

[5] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial Examples: Attacks and Defenses for Deep Learning." arXiv, Jul. 06, 2018. doi: 10.48550/arXiv.1712.07107.

[6] A. Wali et al., "Generative Adversarial Networks for speech processing: A review," Computer Speech & Language, vol. 72, p. 101308, Mar. 2022, doi: 10.1016/j.csl.2021.101308.

[7] Y. Qian, H. Hu, and T. Tan, "Data augmentation using generative adversarial networks for robust speech recognition," Speech Communication, vol. 114, pp. 1–9, Nov. 2019, doi: 10.1016/j.specom.2019.08.006.

[8] D. Saxena and J. Cao. 2021. "Generative Adversarial Networks (GANs): Challenges, Solutions, and Future Directions". ACM Comput. Surv. 54, 3, Article 63 (April 2022), 42 pages. https://doi.org/10.1145/3446374

[9] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 788–798, May 2011, doi: 10.1109/TASL.2010.2064307.

[10] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2014, pp. 1695–1699. doi: 10.1109/ICASSP.2014.6853887.

[11] N. S. Ibrahim and D. A. Ramli, "I-vector Extraction for Speaker Recognition Based on Dimensionality Reduction," Procedia Computer Science, vol. 126, pp. 1534–1540, 2018, doi: 10.1016/j.procs.2018.08.126.

[12] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," Digital Signal Processing, vol. 10, no. 1, pp. 19–41, Jan. 2000, doi: 10.1006/dspr.1999.0361.

[13] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2014, pp. 4052–4056. doi: 10.1109/ICASSP.2014.6854363.

[14] N. Chen, Y. Qian, and K. Yu, "Multi-task learning for text-dependent speaker verification," in Interspeech 2015, ISCA, Sep. 2015, pp. 185–189. doi: 10.21437/Interspeech.2015-81.

[15] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Apr. 2018, pp. 5329–5333. doi: 10.1109/ICASSP.2018.8461375.

[16] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-End Text-Dependent Speaker Verification." arXiv, Sep. 27, 2015. doi: 10.48550/arXiv.1509.08062.

[17] S.-X. Zhang, Z. Chen, Y. Zhao, J. Li, and Y. Gong, "End-to-End Attention based Text-Dependent Speaker Verification." arXiv, Jan. 02, 2017. doi: 10.48550/arXiv.1701.00562.

[18] C. Zhang and K. Koishida, "End-to-End Text-Independent Speaker Verification with Triplet Loss on Short Utterances," in Interspeech 2017, ISCA, Aug. 2017, pp. 1487–1491. doi: 10.21437/Interspeech.2017-1608.

[19] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized End-to-End Loss for Speaker Verification." arXiv, Nov. 09, 2020. doi: 10.48550/arXiv.1710.10467.

[20] Le, N., Odobez, J.-M. (2018) Robust and Discriminative Speaker Embedding via Intra-Class Distance Variance Regularization. Proc. Interspeech 2018, 2257-2261, doi: 10.21437/Interspeech.2018-1685

[21] Bhattacharya, G., Alam, J., Kenny, P. (2019) Deep Speaker Recognition: Modular or Monolithic? Proc. Interspeech 2019, 1143-1147, doi: 10.21437/Interspeech.2019-3146

[22] M. Ravanelli and Y. Bengio, "Learning Speaker Representations with Mutual Information." arXiv, Apr. 05, 2019. doi: 10.48550/arXiv.1812.00271.

[23] M. Razavi et al., "Machine Learning, Deep Learning and Data Preprocessing Techniques for Detection, Prediction, and Monitoring of Stress and Stress-related Mental Disorders: A Scoping Review." arXiv, Aug. 08, 2023. doi: 10.48550/arXiv.2308.04616.

[24] V. Tiwari, "MFCC and its applications in speaker recognition," International journal on emerging technologies, vol. 1, no. 1, pp. 19–22, 2010.

[25] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," in Interspeech 2017, Aug. 2017, pp. 2616–2620. doi: 10.21437/Interspeech.2017-950.

[26] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AISHELL-1: An Open-Source Mandarin Speech Corpus and A Speech Recognition Baseline." arXiv, Sep. 16, 2017. doi: 10.48550/arXiv.1709.05522.

[27] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Apr. 2015, pp. 5206–5210. doi: 10.1109/ICASSP.2015.7178964.

[28] J. Terven, D. M. Cordova-Esparza, A. Ramirez-Pedraza, and E. A. Chavez-Urbiola, "Loss Functions and Metrics in Deep Learning." arXiv, Sep. 06, 2023. doi: 10.48550/arXiv.2307.02694.

[29] T. O. Hodson, T. M. Over, and S. S. Foks, "Mean Squared Error, Deconstructed," Journal of Advances in Modeling Earth Systems, vol. 13, no. 12, p. e2021MS002681, 2021, doi: 10.1029/2021MS002681.