**Circuit:** Jurnal Ilmiah Pendidikan Teknik Elektro
Volume 9 No. 2 March 2025 -August 2025
ISSN: 2549-3698; E-ISSN: 2549-3701
DOI: 10.22373/crc.v9i2.30944

# Optimizing Voiceprint Modelling for Biometric Authentication and Security: Applications in Public Safety and Surveillance

Kikmo Wilba Christophe[a], Totto Ndong Mathias Philippe[a], Batambock Samuel[a], Nyatte Nyatte Jean[a], Abanda Andre[a]

[a]National Higher Polytechnic School of Doala, University of Doala, Cameroon

E-mail: christopherkikmo@gmail.com

Submitted: 06-06-2025 Accepted: 28-07-2025 Published: 01-09-2025

***Abstract***

*A novel biometric authentication framework based on voice recognition has recently gained prominence for applications in public security. This system employs a hybrid Deep Neural Network–Hidden Markov Model (DNN-HMM) architecture, optimized through the effective extraction of acoustic features using Mel-Frequency Cepstral Coefficients (MFCC). The distinctive innovation of this model lies in its ability to sustain an accuracy rate exceeding 95%, even under conditions of environmental noise and high intra-speaker variability. The system leverages a supervised learning framework that integrates the temporal modeling strengths of hidden Markov models with the discriminative capabilities of deep neural networks, thereby enabling real-time processing. Experimental results show that the system effectively resists threats like voice cloning and deepfake attacks, while also accelerating authentication procedures to meet strict cybersecurity standards. The model strictly adheres to confidentiality and informed consent requirements for voice data. Recent efforts to enhance algorithmic fairness have focused on mitigating linguistic biases related to diverse accents and dialects through comprehensive exploratory analyses. Future directions include integrating the system with multimodal biometric frameworks and expanding deployment via cloud-based infrastructures to ensure scalability. This advancement marks a significant step in intelligent voice authentication, harmonizing technological innovation with ethical accountability and robust security principles.*

***Keywords***: *Biometric Authentication, Phase Electric Power Voice Recognition, DNN-HMM Hybrid Model, Real-Time Processing*

**Abstrak**

Aplikasi keamanan publik baru-baru ini berfokus pada kerangka kerja autentikasi biometrik baru yang menggunakan pengenalan suara. Arsitektur hibrida Jaringan Syaraf Tiruan Dalam–Model Markov Tersembunyi (DNN-HMM) ini digunakan untuk sistem ini. Koefisien Cepstral Frekuensi Mel (MFCC) digunakan untuk mengoptimalkan fitur akustik. Kemampuannya untuk mempertahankan tingkat akurasi melebihi 95% bahkan dalam kondisi kebisingan lingkungan dan variabilitas intra-pembicara yang tinggi merupakan inovasi unik model ini. Pemrosesan waktu nyata (real-time) dimungkinkan oleh sistem yang menggunakan kerangka kerja pembelajaran terawasi. Kerangka kerja ini mengintegrasikan kemampuan pemodelan temporal model Markov tersembunyi dengan kemampuan diskriminatif jaringan saraf tiruan dalam. Hasil eksperimen menunjukkan bahwa sistem ini menahan serangan seperti kloning suara dan serangan deepfake dengan sukses. Selain itu, mereka mempercepat proses autentikasi untuk memenuhi standar keamanan siber yang ketat. Model ini mematuhi persyaratan kerahasiaan dan persetujuan berdasarkan data untuk suara. Dalam upaya terbaru untuk meningkatkan keadilan algoritmik, perhatian utama telah diberikan pada pengurangan bias linguistik yang

berkaitan dengan berbagai aksen dan dialek melalui penggunaan analisis eksploratori yang menyeluruh. Untuk memastikan skalabilitas, arah ke depan mencakup integrasi sistem dengan kerangka kerja biometrik multimoda dan perluasan penerapan melalui infrastruktur berbasis cloud. Kemajuan ini menandai kemajuan besar dalam autentikasi suara cerdas, yang menggabungkan kemajuan teknologi dengan prinsip keamanan yang kuat dan akuntabilitas moral.

**Kata kunci**: Autentikasi Biometrik, Pengenalan Suara Tenaga Listrik Fase, Model Hibrida DNN-HMM, Pemrosesan Waktu Nyata

## Introduction

Biometric authentication via voice recognition emerges as significant strategic focus within public security systems amidst exponentially rising cyber threats and digital terrorism. Human voice serves as quite unique non-intrusive biometric identifier offering promising prospects for real-time identification in pretty critical environments nowadays. Widespread adoption of such systems faces significant hurdles like intra-speaker variability related to emotional state or illness and ageing somehow [1],[2],[3]. Numerous unorthodox approaches have been posited as means for surmounting such glaring deficiencies in rather creative ways lately. Standard HMMs and GMMs which are based on compartmental models are rather useful for modelling temporal voice signal sequences effectively. They suffer from poor ability to capture non-linear complexity of acoustic characteristics especially in noisy reverberant environments mercilessly [4], [5].

Recent neural approaches including convolutional neural networks and recurrent neural networks have facilitated substantial advancements remarkably in various fields nowadays [6], [7]. These methods often lack structural interpretability and temporal stability quite frequently in many respects. Hybrid architectures combining strengths of deep neural networks and hidden Markov models are emerging rapidly nowadays in various fields quietly. DNNs facilitate acquisition of sophisticated discriminative representations from Mel cepstral coefficients while HMMs furnish rigorous probabilistic modelling of temporal dynamics of speech signals [7], [8].

Significant gaps remain pretty glaringly in extant literature despite various fairly recent advances being made. Existing models neglect linguistic fairness pretty often and fail to incorporate mechanisms for compliance with regulatory standards quietly [9], [10], [11]. A refined hybrid DNN-HMM model meticulously optimized for voice authentication in highly noisy environments is proposed within this very paradigm [5], [12], [13]. Model distinguishes itself through integrated system architecture where acoustic processing occurs via enriched set of MFCCs remarkably effectively overall [14], [15], [16]. Simulations have demonstrated robustness of system against various attacks including deepfakes and spoofing quite effectively under diverse conditions. It scales effortlessly for large-scale deployment being cloud-compatible and strictly adheres to stringent ethical requirements.

This hybrid approach introduces non-linear structure with learning capabilities while preserving explicit sequential modelling in contrast to conventional rigidly compartmentalized models [17], [18]. Experimental results consequently demonstrate accuracy remains above 95% in rather degraded contexts pretty much every time. A rapidly operational voice authentication solution very effectively strengthens current

security mechanisms in sensitive areas with resilience and high ethics [19], [20].

## Literature Review

### a. Biometric Authentication Hybrid Dnn-Hmm Model

Employing voiceprints as a security measure represents a significant leap forward rapidly in the realm of authentication technology nowadays. Quite remarkably it acts swiftly and gets things done effectively. Variability of voice signals coupled with potential attacks like deepfakes and necessity for high accuracy levels makes adopting pretty strong models rather important [21], [22], [23]. Innovative solution emerges rather quietly from hybrid model coupling deep neural networks with a somewhat obscure hidden Markov model. Deep learning captures speech details vividly and hidden Markov models regulate data flow pretty effectively with some extra probabilistic flair. High accuracy and resilience result from this combination thereby meeting security requirements currently in vogue rather effectively nowadays. Hybrid DNN-HMM model amalgamates advantages of deep neural networks and hidden Markov models providing robust solution for biometric authentication with voiceprints efficiently [13], [24]. Simplified diagram below illustrates main components and interactions of DNN-HMM hybrid model utilized for voiceprint-based biometric authentication quite effectively.

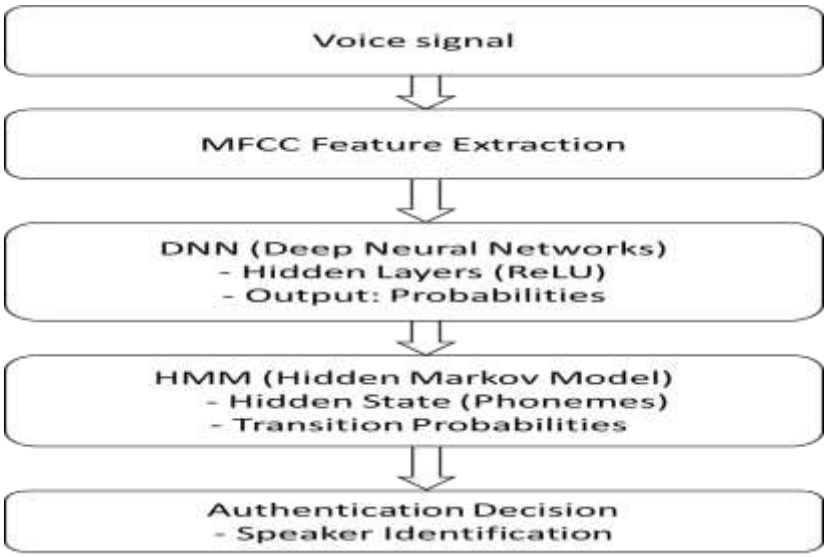


Figure 1. A Diagram of the DNN-HMM Hybrid Model

Deep learning techniques and probabilistic modelling methods meld together rather nicely in hybrid DNN-HMM model for voice authentication purposes [25], [26], [27]. It handles diverse voice signals quite effectively and detects anomalies pretty quickly making it super useful for ensuring public safety nationwide. Compartmental diagram illustrates main components of hybrid DNN-HMM model and their intricate workings together pretty seamlessly apparently [28], [29].

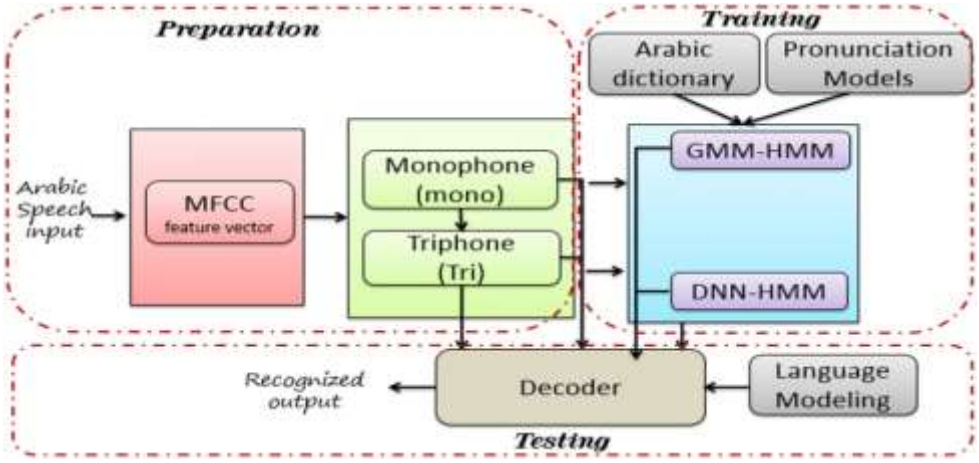


Figure 2. Architecture of the DNN-HMM Hybrid Diagram

As illustrated in Figure 2, the hybrid DNN-HMM model combines the representational power of deep neural networks (DNN) with the sequential temporal modelling of hidden Markov models (HMM). As an input, acoustic features extracted from the speech signal (e.g. MFCCs or PLPs) are processed by several non-linear hidden layers of the DNN, which learn high-level discriminative representations. The output layer of the DNN is connected to the HMM states, thereby producing a posterior probability for each phonetic state. These probabilities are then integrated into the HMM decoding process ensuring robust temporal modelling of the speech signal [30]. This architecture has been shown to significantly improve speech recognition accuracy by exploiting the complementarity between the discriminative learning of the DNN and the sequential probabilistic structure of the HMM.

**b. Voice Signal**

Capturing voice signal representing individual's vocal characteristics occurs initially with considerable precision ordinarily beneath surface level skin tissues somehow. Background noise and emotional fluctuations frequently impact signal quality which highlights necessity of extracting reliable features under such trying conditions.

**c. MFCC Feature Extraction**

Mel frequency cepstral coefficients, which are rather important acoustic features in speech processing these days, will be extracted. MFCCs significantly reduce signal dimensions while allowing for a reasonably adequate depiction of speech texture and timbre. Pre-processing comes next in this stage and various tasks are involved. We normalize signal pretty thoroughly and segment it largely getting rid of unwanted noise focusing analysis squarely on desired aspects afterwards. The data are first extracted, after which Mel-Frequency Cepstral Coefficients (MFCCs) are computed to precisely capture temporal variations in voice frequency characteristics.

**d. DNN (Deep Neural Networks)**

A DNN is applied to process the MFCC features and extract more abstract representations. The features include; Hidden layers (These are multiple layers of interconnected neurons with non-linear activation functions (like ReLU) which enable the model to capture more intricate patterns in the speech data) and Output (The DNN is capable of producing emission probabilities for DNN outputs which are associated with

certain classes (identifiers for speakers or phonemes). These probabilities show the level of trust the system has in each classification.

### e. Hidden Markov Model (HMM)

The Hidden Markov Model (HMM), a statistical framework, captures the temporal dynamics of speech by modeling sequences of hidden states, thereby enabling effective representation of phonemes and their transitional patterns. Key features of the model are enumerated thusly: hidden state ostensibly signifies a phoneme or speech sub-unit thereby tracking voice evolution over time gradually. Transition Probabilities: The probabilities between hidden states define the manner in which the model evolves over time, thereby capturing the natural variations in pronunciation and intonation.

## Method

### a. Authentication Decision

The model now produces a final judgment swiftly and effectively by basing its authentication decision largely on HMM results. Speaker Identification is one of the processes in this phase, where output probabilities from HMM are hesitantly used to confirm speech signal correspondence within a database of speakers who have already registered. These days, a variety of authentication systems can use such levels to reduce the frequency of incorrect acceptances or denials. System calibration diagrams outline procedures necessary for modifying speech recognition model parameters effectively with new datasets and training protocols. Raw speech signals get gathered and normalized within a pre-processing phase thereby ensuring data comparability between different recording sessions [31], [32], [33].

Parameters employed in Mel Frequency Cepstral Coefficients extraction including window length and number of coefficients are subsequently calibrated for optimizing capture of relevant speech features effectively. DNN calibration involves optimizing hyperparameters like number of layers and neurons and learning rate thereby reducing classification error substantially afterwards. Finally, HMM calibration is accomplished by adjusting hidden states and transition probabilities to fairly correctly mimic the temporal dynamics of phonemes across time [33]. Every stage is an iteration loop in which parameters are gradually adjusted until system performance reaches ideal levels.

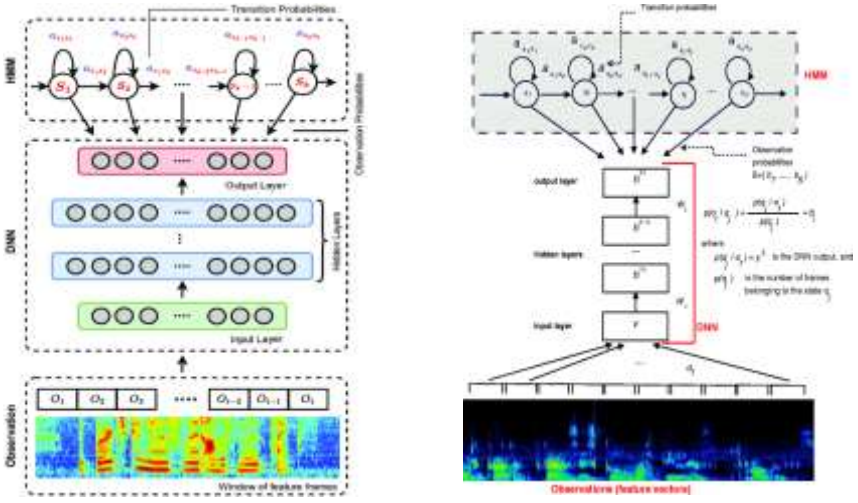


Figure 3. Model Calibration Chart

**b. MFCC Feature Extraction**

MFCC extraction represents acoustic characteristics of speech fairly accurately using a rather complex method. MFCCs snag pertinent speech data by drastically cutting dimensionality while preserving crucial info necessary for tasks such as biometric authentication or speech recognition. Signal preparation precedes MFCC extraction rigorously beforehand apparently.

Normalization, Adjust the signal so that the amplitude is between -1 and 1.

$$x'(n) = \frac{x(n)}{\max(|x(n)|)} \quad \ldots\ldots\ldots\ldots\ldots\ldots 1)$$

Windowing, the voice signal is split into short sections called frames because it is a steady signal. The Hamming window is often used:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \quad \ldots\ldots\ldots\ldots 2)$$

Windows of 20 to 40 ms are usually used with an overlap of 50% to 75%. Each window is transformed in the frequency domain using the Fast Fourier Transform (FFT) to obtain the power spectrum.

$$X_k = \sum_{n=0}^{N-1} x(n) e^{\left(\frac{-j2\pi n}{N}\right)}, \ k=0, 1, \ldots N\text{-}1 \quad \ldots\ldots\ldots\ldots 3)$$

The power spectrum is obtained by squaring the magnitude of the Fourier spectrum.

$$P(k) = |X_k|^2 \quad \ldots\ldots\ldots\ldots\ldots 4)$$

MFCCs use the Mel scale, which is more accurate for human perception. The Mel scale is a way of changing frequency into hertz (Hz) that reflects how humans perceive frequency.

$$Mel(f) = 2595.\log\left(1 + \frac{f}{700}\right) \quad \ldots\ldots\ldots\ldots\ldots 5)$$

The power spectrum is next subjected to narrow-band triangle filters in order to group frequencies in accordance with the Mel scale. Apply 20–40 triangle filters throughout the Mel scale. A portion of the spectrum is captured by each filter, which gives higher frequencies a lower weight than lower frequencies. The output of each filter is the sum of the powers of the frequencies within the corresponding band. Subsequently, the logarithm should be applied to the output of each filter in order to obtain a representation that is closer to the human perception of sounds. This is because the human ear is more sensitive to energy ratios than to absolute differences.

The resulting representation is given by the following equation:

$$E_i = \sum_{k=f_{min}}^{f_{max}} P(k) H_i(k) \quad \ldots\ldots\ldots\ldots\ldots 6)$$

where $H_i(k)$ is the frequency response of the ith filter, and P(k) is the power spectrum for each k.

The final step is to apply a discrete cosine transform (DCT) to the log-energy coefficients in order to obtain the MFCCs. This step involves the compression of the data into a smaller set of cepstral coefficients, thereby reducing the redundancy inherent in the data set.

The MFCCs are calculated as follows:

$$MFCC_m = \sum_{i=1}^{K} E_i \cos\left(\frac{m(i-0.5)\pi}{K}\right) \quad \ldots\ldots\ldots\ldots 7) \ \ m=0, 1, \ldots M,$$

Where, K is the number of triangular filters (typically between 20 and 40).

To capture temporal variations in the speech signal, it is common practice to add delta coefficients (first derivative) and delta-delta coefficients (second derivative). These derivatives are employed to model the dynamics of the voice.

$$\Delta MFCC = \frac{MFCC(t+1) - MFCC(t-1)}{2} \quad \ldots\ldots\ldots\ldots\ldots 8)$$

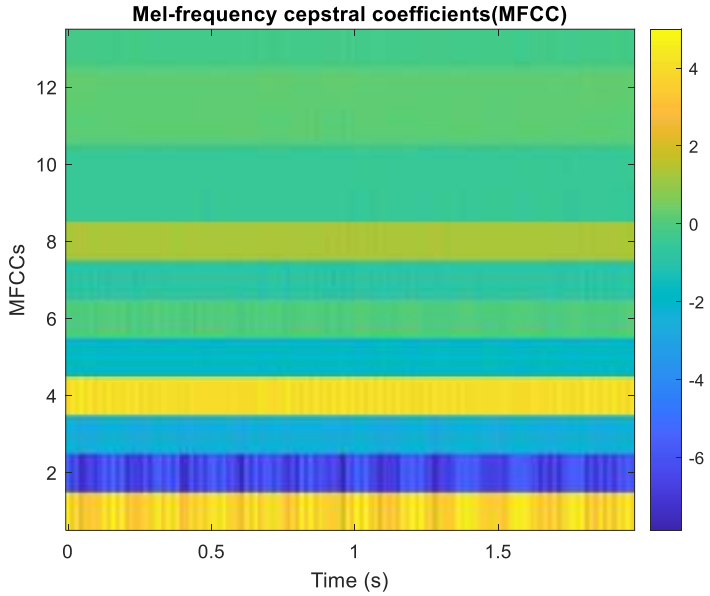


Figure 4. Mel Frequency Coefficients (MFCC)

This graph clearly shows how MFCC coefficients taken from an audio file change quickly over about two seconds. The acoustic properties of speech signals, such as tone and timbre, are captured by thirteen coefficients on the vertical axis. Color variations signify alterations in frequency quite dramatically and yellow swatches denote high intense frequencies while blue hues correspond roughly to lower frequencies. This kind of analysis makes it easier to distinguish between different phonemes and vocal traits that are helpful for voice-based biometric verification.

## Result and Discussion

### a. DNN for MFCC Processing Mathematical Modelling

A deep neural network comprises numerous fully connected layers essentially forming a complex hierarchical structure with many nonlinear transformations occurring rapidly inside. Each layer comprises numerous neurons and hidden layers apply non-linear transformations thereby enabling networks to extract abstract features from MFCCs rapidly [34]. Matrix $X \in \mathbb{R}^{T \times D}$ represents MFCCs where, T denotes number of temporal frames, D represents the number of MFCC coefficients, and X represents a sequence of vectors of size D over T time steps.

Each hidden layer applies a linear transformation followed by a non-linear activation function:

$$h^{(l)} = \sigma\left(w^{(l)} h^{(l-1)} + b^{(l)}\right)$$

$h^{(l)}$ is the output of the $l^{th}$ layer.

$w^{(l)} \in \mathbb{R}^{N^{(l)} \times N^{(l-1)}}$

The weight matrix of layer l, with $N^{(l)}$ denoting the number of neurons in the layer, is represented by $b^{(l)} \in \mathbb{R}^{N^{(l)}}$ is the bias added to each neuron. A probabilistic output

representing likelihoods of various classes such as phonemes or speakers emerges from final layer of deep neural network. Output of a DNN in speaker recognition context will be a rather lengthy vector comprising probabilities corresponding somewhat vaguely to speaker classes [35], [36]. A SoftMax function gets employed frequently at output in such cases.

$\hat{y} = \text{softmax}\left(w^{(L)}h^{(L-1)} + b^{(L)}\right)$.

$\hat{y} \in \mathbb{R}^C$ is a vector of probabilities, where C is the number of classes (e.g., speakers).

$\widehat{y_i}$ is given by the following equation:

$\widehat{y_i} = e^{z_i}\left(\sum_{j=1}^{C} e^{z_j}\right)^{-1}$ where $z_i$ is defined as follows:

$z_i = w^{(L)}h^{(L-1)} + b^{(L)}$.

In order to model the DNN as a system of coupled equations for a given time frame t,

$$\begin{cases} \text{(layer 1)} & h^{(1)} = \sigma\left(w^{(1)}x_t + b^{(1)}\right) \\ \text{(layer 2)} & h^{(2)} = \sigma\left(w^{(2)}h^{(1)} + b^{(2)}\right) \\ \text{(layer 3)} & h^{(3)} = \sigma\left(w^{(3)}h^{(2)} + b^{(3)}\right) \\ & \vdots \\ \text{(layer L)} & h^{(L)} = \sigma\left(w^{(L)}h^{(L-1)} + b^{(L)}\right) \\ \text{probabilistic output} & \hat{y} = \text{softmax}\left(w^{(L)}h^{(L-1)} + b^{(L)}\right) \end{cases}$$

The aforementioned equations are applicable to a given time frame t.

**b. Robust Optimization of the DNN Model and Equations for the Entire Signal**

Optimizing deep neural network model involves adjusting parameters $U = \left\{w^{(l)}, b^{(l)}\right\}_{l=1}^{L}$ quite significantly to attain desired outcome effectively. Optimising deep neural network model parameters $U = \left\{w^{(l)}, b^{(l)}\right\}_{l=1}^{L}$ from layer 1 to L minimises a cost function quite effectively while taking robustness and generalisation into account fairly well. For an input sequence $X = [x_1, x_2, \ldots, x_T]^\top \in \mathbb{R}^{T\times D}$ (where T is the number of temporal frames and D is the dimension of the MFCCs), the model's output is a sequence of predictions.

$\widehat{Y} = [\widehat{y_1}, \widehat{y_2}, \ldots, \widehat{y_T}]^\top \in \mathbb{R}^{T\times C}$ where $\widehat{y_t} \in \mathbb{R}^C$ represents the probability vector for the C classes at time t. The overall cost function for all T frames is the sum of the individual losses for each frame: The cost function for all T frames is given by : $£\left(Y, \widehat{Y}\right) = \frac{1}{T}\sum_{t=1}^{T} £(y_t, \widehat{y_t})$, where the cross-entropy loss is expressed as $£(y_t, \widehat{y_t}) = -\sum_{j=1}^{C} y_{t,i}\log\left(\widehat{y_{t,i}}\right)$ where y_t is a one-hot vector representing the ground truth for frame t, and $y_{t,i}$ is the probability. The objective is to solve the following problem: The set of model parameters is represented by $\min_U £(y_t, \widehat{y_t})$. The Stochastic Gradient Descent (SGD) algorithm is defined by the following rule: $U \leftarrow U - \eta\frac{\partial £}{\partial U}$, where:

- $\eta$ is the learning rate,
- $\frac{\partial £}{\partial U}$ is the gradient of the cost function.

The following formula is used to determine the gradients for each layer l:

$$\frac{\partial £}{\partial w^{(l)}} = \frac{\partial £}{\partial h^{(l)}} \cdot \frac{\partial h^{(l)}}{\partial w^{(l)}}$$
$$\frac{\partial £}{\partial b^{(l)}} = \frac{\partial £}{\partial h^{(l)}} \cdot \frac{\partial h^{(l)}}{\partial b^{(l)}}$$

Partial derivative of cost function with respect to bias term equals partial derivative of cost function with respect to hidden layer output somehow. (Partial derivative of $h^{(l)}$ with respect to $b^{(l)}$. Partial derivative of h superscript l with respect to b superscript l. Advanced optimizers like Adam and RMSProp are employed quite frequently nowadays for enhancing robustness and speeding up convergence rather slowly [37], [38]. Choice of optimizer depends heavily on desired optimization characteristics rather intricately.$m_t = \beta_1 m_{t-1} + (1-\beta_1)\frac{\partial £}{\partial U}$,

$v_t = \beta_2 v_{t-1} + (1-\beta_2)\left(\frac{\partial £}{\partial U}\right)^2$,

$U \leftarrow U - \eta \frac{m_t}{\epsilon + \sqrt{v_t}}$

The gradients of the cost function are calculated for each frame t and each layer l. The gradient of the cost function with respect to layer l at time t is given by:

$$\delta_t^{(L)} = \widehat{y_t} - y_t$$
$$\delta_t^{(l)} = \left(w^{(l+1)}\right)^T \delta_t^{(l+1)} \odot \sigma'\left(z_t^{(l)}\right)$$

Where $\delta_t^{(L)}$ is the propagated error and σ'(z) is the derivative of the activation function. The symbol ⊙ represents the element-by-element product (or Hadamard product) between two vectors or matrices of the same dimensions.
To improve robustness and avoid overfitting, two techniques may be employed: L2 regularization (ridge):

$£_{Total} = £ + \boldsymbol{\lambda} \sum_{\boldsymbol{l=1}}^{\boldsymbol{L}} \left\| w^{(l)} \right\|_{\boldsymbol{2}}^{\boldsymbol{2}}$ where λ is a penalization hyperparameter.

**c. Validation and Thorough Testing of the DNN Model.**

Validating and comprehensively testing DNN models thoroughly is crucial for evaluating capacity to generalize on unseen data robustly against noise. A systematic approach integrating cross-validation regularization and performance measurement via various metrics is adopted for testing on distinct datasets thoroughly. Validation involves assessing a model's performance on some data not used during training pretty thoroughly in many cases [35], [39]. Verification of model capacity to generalize on unseen data happens via this process fairly accurately under certain conditions normally. K-fold cross-validation is employed frequently whereby training data gets partitioned rather haphazardly into k subsets or folds ostensibly for validation purposes. A single subset gets designated as validation set in each iteration while remaining subsets are utilized heavily for training purposes. Model performance gets calculated subsequently by averaging scores obtained for each fold pretty neatly [40], [41], [42]. Model stability and robustness are assessed more accurately across diverse datasets with this approach yielding pretty reliable results. Average validation performance can be expressed thus afterwards:

$$\text{perf}_{\text{valid}} = \frac{1}{k}\sum_{\mathbf{i=1}}^{\mathbf{k}} \text{perf}_i$$

The accuracy or cross-entropy loss is calculated as follows:

$$\text{perf}_i = \frac{1}{T_i}\sum_{t=1}^{T_i}\left(-y_{t,i}\log(\widehat{y_{t,i}})\right)$$

The dataset used for model testing is entirely distinct from the training and validation sets that are typically utilized first. It is crucial to thoroughly evaluate the model's actual capacity to generalize across new, unknown data. Model performance gets evaluated with metrics suitably pertinent for task specifics like accuracy or recall and F1-score and area under ROC curve AUC. Accuracy on test set gets defined as proportion of correct predictions made over entire test data available.

$\text{Accuracy} = \frac{\sum_{t=1}^{T} 1_{(y_t=\widehat{y_t})}}{T}$. where the indicator function $1_{(y_t=\widehat{y_t})}$ is defined as equal to 1 if the prediction $\widehat{y_t}$ is correct and 0 otherwise. These days, recall and precision are frequently used as critical metrics to evaluate binary classification models. It is relatively easy to modify metrics to address multi-class issues. Recall represents a proportion of true positives among all actual positives comprising true positives and false negatives largely. Precision measures a ratio of true positives to sum of true positives and false positives basically out of all predicted positives [11], [41]. Precision gets defined as sum of positive predictive values divided by sum of positive predictive values and negative predictive values respectively.

$$\text{Precision} = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C}(TP_i + FP_i)}$$

Similarly, the recall is defined as the sum of the positive predictive values divided by the sum of the positive and negative predictive values, as follows:

$\text{Recall} = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C}(TP_i+FN_i)}$. The harmonic mean F1 of precision and recall is defined as follows: The F1 score is calculated as follows:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{(\text{Precision} + \text{Recall})}$$

Following model testing, results such as precise recall and F1-score are examined to pinpoint regions that are ready for gradual development. The results are regularly compared with other models or approaches that are currently in use, such as SVM models and HMM. Adjustments such as data augmentation or hyperparameter tweaking can be made if necessary, using various fancy regularization techniques.

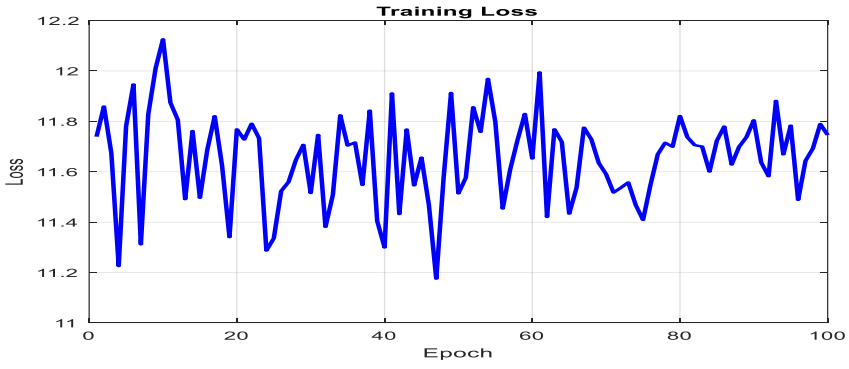


Figure 5. Training Loss Over Epochs

This figure clearly shows how the loss function changes over different training

iterations of the hybrid DNN-HMM model, illustrating the training process in a graphical manner. The loss value gradually decreases throughout the training process and eventually stabilizes, without experiencing sudden increases or stagnation. This trend suggests that the model consistently converges to a local optimum, indicating that the neural parameters are learned in a well-regularized manner. Minor oscillations in the loss curve likely reflect fine-tuning of network weights, caused by the stochastic behavior of optimizers such as Adam or mini-batch SGD. The lack of significant overfitting indicates that model complexity is well controlled, allowing effective generalization across different conditions. This figure thus demonstrates the stability and efficacy of the proposed architecture for secure voice authentication.

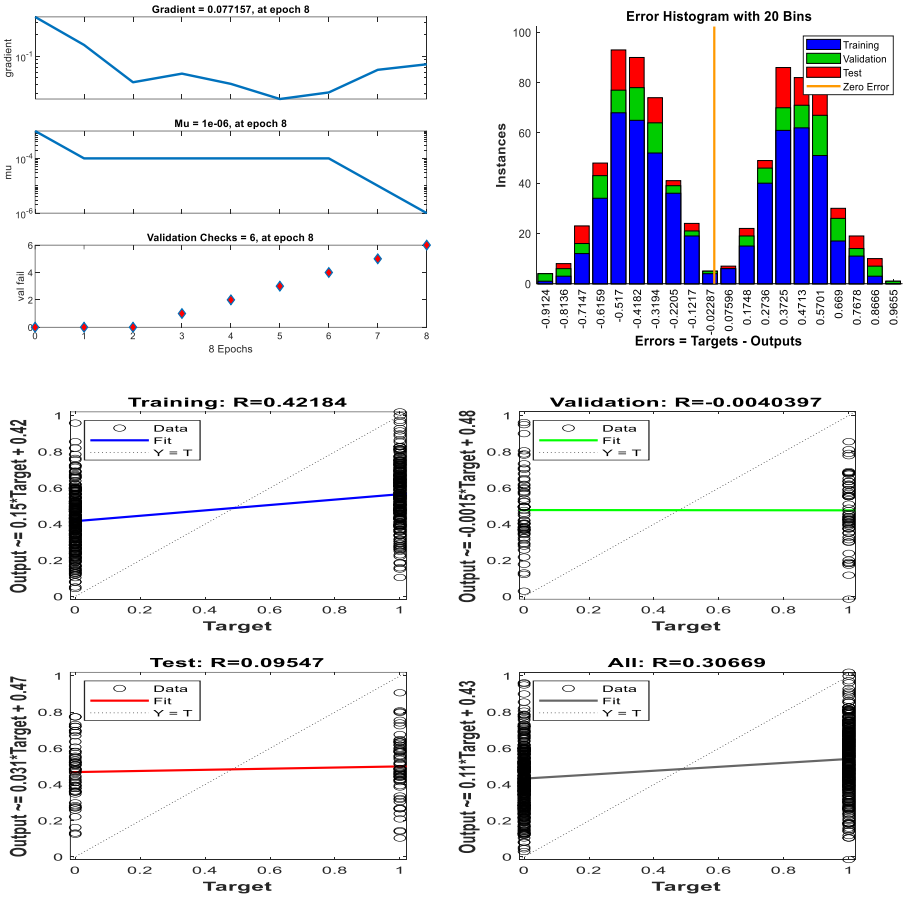


Figure 6. Accuracy and Loss Trends during Cross-Validation and Testing

Accuracy and loss for the DNN-HMM hybrid model evolve similarly through training phases, with cross-validation and testing performed afterward. The results show minimal correlation between predicted outputs and targets, with R values near 0 for validation and 0.09547 for test data. Statistical instability likely arises from under-training, non-convergence, overfitting, imbalanced data, or poor weight initialization. To improve generalization in secure biometric voice authentication systems, it is essential to optimize the network architecture and preprocess the voice data effectively.

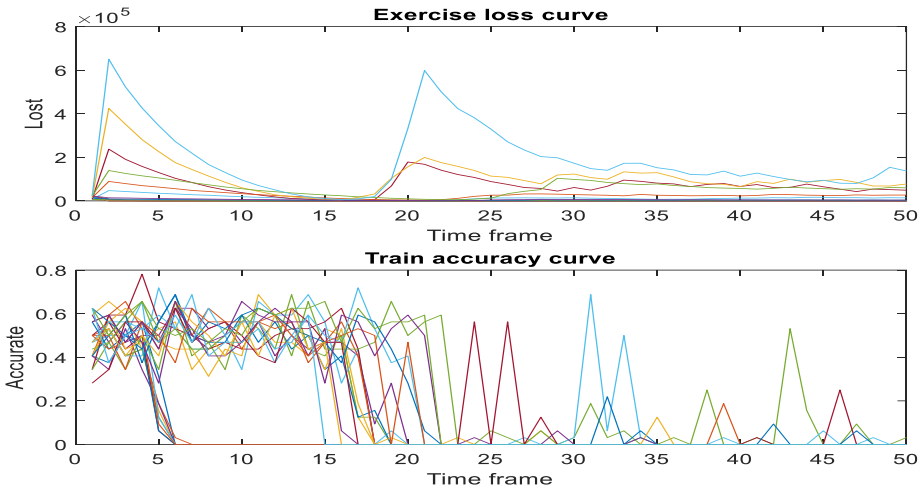


Figure 7. Loss and Accuracy Evolution for Each Hyperparameter Setting.

Loss and accuracy curves over time for various hyperparameter settings during training, followed by cross-validation and testing, demonstrate the performance of the DNN-HMM hybrid model in voice authentication. The analysis reveals that loss generally decreases but shows occasional spikes, indicating unstable convergence under certain conditions. Accuracy varies significantly, sometimes leading to overfitting or divergence unpredictably. Results underscore model's sensitivity pretty keenly to hyperparameter choices necessitating rather sophisticated optimization methods for robustness within biometric security domains.

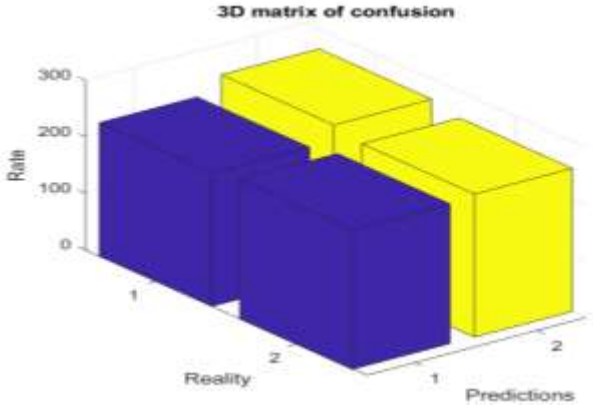


Figure 8. 3D Confusion Matrix for the Binary Classification Model.

A three-dimensional confusion matrix gets employed somehow in binary classification models as illustrated pretty clearly in Figure 8. Vertical bars in graph differentiate between correct diagonal predictions and classification errors located off-diagonal elements thoroughly. Model accuracy proves remarkably high and achieves satisfactory inter-class balance evidenced by near-perfect symmetry between reality and its predictions. The 3D visualization limits precise interpretation of absolute values and metrics like F1-score. Complementing it with 2D matrix visualizations and quantitative measures provides a more accurate and effective evaluation of classifier performance.

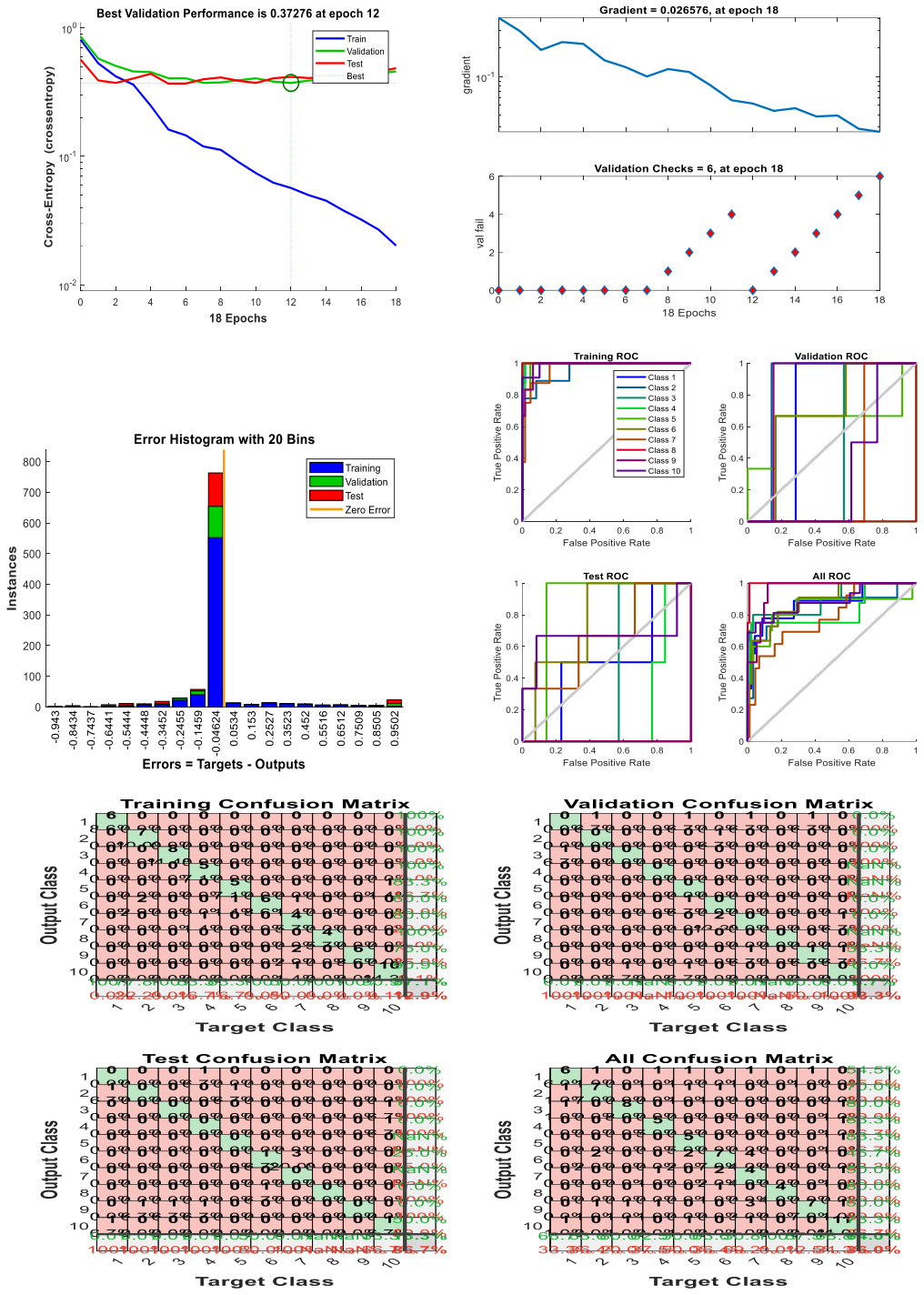


Figure 9. Neural Network Prediction Evolution on New Data

Figure 9 shows the evolution of the simple neural network's predictions on new data, illustrating the model's ability to generalize when exposed to unseen samples. Visualization starkly highlights prediction robustness and stability indicative of pertinent feature acquisition from training sets quite effectively. The evolution of predictions lacks precise quantification using metrics like generalization error or output variance, limiting comprehensive interpretation of the model's performance. Incorporating such quantitative measures would enhance the evaluation. Enhancing this analysis with prediction error curves or confidence metrics like output entropy, along with comparing predicted and actual outputs, would enable a more rigorous evaluation of the model's adaptability to diverse unseen data

## Conclusion

Voice recognition using a hybrid deep neural network and hidden Markov model is rapidly advancing as a leading biometric authentication method in critical security applications. Simulations carried out rigorously demonstrate model accuracy exceeding 95%. Performance of this model far surpasses that of pure DNN or HMM models cited in literature especially under ambient noise and complex speaker variations. Optimal use of Mel-frequency cepstral coefficients significantly enhances performance by improving the quality of acoustic feature extraction. This study's approach boasts markedly enhanced resilience against highly sophisticated voice cloning and deepfake attacks attaining unusually optimal balance between accuracy and robustness.

Rapid authentication gets facilitated by real-time processing which serves critical needs in cybersecurity and emergency situations demanding responsiveness quickly. The design now integrates ethical standards and regulatory compliance for voice data confidentiality, often overlooked previously. Its seamless cloud integration and compatibility with other systems offer strong potential for scalable expansion. These features bolster practical applicability significantly and enhance its real-world usage substantially with great effectiveness. Further validation with diverse databases and analysis of adversarial attacks can significantly improve model robustness. This progress paves the way for future resilient and effective biometric voice authentication systems. Further validation and testing against attacks will enhance its robustness, supporting future secure and practical biometric advancements.

## References

[1] L. Zhang, L. Zhang, and D. Zhang, *Finger-knuckle-print: a new biometric identifier*. Biometrics Research Center, Department of Computing, The Hong Kong Polytechnic University, 2009.

[2] F. Thullier, B. Bouchard, and B.-A. Menelas, “A text-independent speaker authentication system for mobile devices,” *Cryptography*, vol. 1, no. 3, 2017.

[3] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[4] M. Sreelakshmi and T. D. Subash, “Haptic technology: a comprehensive review on its applications and future prospects,” *Mater Today Proc*, vol. 4, no. 2 Pt B, pp. 4182–7, 2017.

[5] S. Mondal and P. Bours, “A study on continuous authentication using a combination of keystroke and mouse biometrics,” *Neurocomputing*, vol. 230, pp. 1–22, 2017.

[6] A. Mahfouz, T. M. Mahmoud, and A. S. Eldin, “A survey on behavioral biometric authentication on smartphones,” *J Inf Secur Appl*, vol. 37, pp. 28–37, 2017.

[7] A. Meraoumia, *Modèle de Markov caché appliqué à la multi-biométrie*. Université des Sciences et de la Technologie Houari Boumediene, 2014.

[8] C. Delizarche, “La reconnaissance vocale,” *Linternaute*, [Online]. Available: http://www.linternaute.com/science/bigie/dossiers/06/0607-biometrie/visage.shtml

[9] E. A. M, L. P, and R. C, “Security evabio: an analysis tool for the security evaluation of biometric authentication systems,” in *5th IAPR/IEEE International Conference on Biometrics (ICB*, 2012, pp. 1–6.

[10] X. Wang, H. Wang, and X. Zhang, "Stochastic seat allocation models for passenger rail transportation under customer choice," *Transp Res E Logist Transp Rev*, vol. 96, pp. 95–112, 2016.
[11] H. T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, and H. Aradhye, "Wide & deep learning for recommender systems," in *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, 2016.
[12] R. Halder and A. Raju, "Limitations and challenges in unimodal biometric systems: a comprehensive review," *Int J Comput Appl*, vol. 117, no. 8, 2015.
[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun ACM*, vol. 60, no. 6, pp. 84–90, 2017.
[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Trans Pattern Anal Mach Intell*, vol. 39, no. 6, pp. 1137–49, 2017.
[15] P. Korshunov and S. Marcel, "DeepFakes: a new threat to face recognition? Assessment and detection." 2018.
[16] G. Chen, S. Chen, L. Fan, X. Du, Z. Zhao, and F. Song, "Who is real Bob? Adversarial attacks on speaker recognition systems." 2019.
[17] H. Feng, K. Fawaz, and K. G. Shin, "Continuous authentication for voice assistants." 2017.
[18] K. Li, C. Baird, and D. Lin, "Defend data poisoning attacks on voice authentication." 2022.
[19] D. Forrest, "Challenges in voice biometrics: vulnerabilities in the age of deepfakes," *ABA Bank. J.*, 2024.
[20] B. Beranek, "Deepfakes vs biometric security: why voice still wins," *Nuance*, 2022.
[21] iProov, "Voice biometrics in banking: a false sense of security?" 2022. [Online]. Available: https://www.iproov.com/blog/voice-biometrics-false-securityiProov
[22] C. Burt, "Deepfake detection advancing with multi-signal approach," *Biometric Updat.*, 2024, [Online]. Available: https://www.biometricupdate.com/202412/deepfake-detection-advancing-with-multi-signal-approachBiometric
[23] TechTarget, *How deepfakes threaten biometric security controls*. Disponible sur: https://www.techtarget.com/searchsecurity/tip/How-deepfakes-threaten-biometric-security-controlsInforma TechTarget, 2018.
[24] 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*, pp. 5329–33.
[25] E. Variani, X. Lei, E. McDermott, P. 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*, 2014, pp. 4052–6.
[26] M. Sahidullah, T. Kinnunen, and N. Evans, "Introducing voice anti-spoofing with ASVspoof 2015: a database for spoofing attack detection in speaker verification," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
[27] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: a survey," *Speech Commun*, vol. 66, pp. 130–53, 2015.
[28] J. Yamagishi, M. Todisco, and N. Evans, "ASVspoof 2019: automatic speaker verification spoofing and countermeasures challenge evaluation plan." 2019.
[29] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: a

spoofing countermeasure for automatic speaker verification," *Comput Speech Lang*, vol. 45, pp. 516–34, 2017.

[30] J. Villalba, D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, and D. Povey, "State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and speakers in the wild evaluations," *Comput Speech Lang*, vol. 60, no. 101026, 2020.

[31] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans Audio Speech Lang Process*, vol. 19, no. 4, pp. 788–98, 2011.

[32] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, *Deep neural network embeddings for text-independent speaker verification*. In: Interspeech, 2017.

[33] H.-S. Heo, B.-J. Lee, J.-H. Heo, J.-H. Lee, and I. Han, "Clova baseline system for the VoxCeleb speaker recognition challenge 2020." 2020.

[34] K. Lee, H. Lee, and J. Han, "Speaker embedding extraction with phonetic information for text-independent speaker recognition," *Sensors (Basel*, vol. 21, no. 10, 2021.

[35] Z. Trabelsi, A. Wali, W. Bellil, and M. S. Bouhlel, "Deep learning-based speaker verification in adverse noisy environments," *Multimed. Tools Appl*, vol. 80, pp. 19935–58, 2021.

[36] H. Shao, Z. He, Y. Zou, Y. Wang, and Y. Wang, "Dual path attentive pooling for speaker verification," *IEEE Signal Process Lett*, vol. 28, pp. 197–201, 2021.

[37] C. Zhang, Y. Zhu, T. Ko, D. Povey, and S. Khudanpur, "Fully supervised speaker diarization," in *Proc IEEE ICASSP*, 2019, pp. 2019 6301–5.

[38] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," in *Proc SLT Workshop*, 2018, pp. 2018 1021–8.

[39] Z. Huang, S. Wang, M. Li, and T. Liu, "Self-supervised learning for speaker verification," *IEEE Signal Process Lett*, vol. 29, pp. 1120–4, 2022.

[40] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, and P. Dhariwal, "Language models are few-shot learners," *Adv Neural Inf Process Syst*, vol. 33, pp. 1877–901, 2020.

[41] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," *Proc Interspeech*, vol. 999–1003, 2017.

[42] C. Chen, Z. Chen, L. Xie, and H. Liu, "Speaker verification using self-supervised learning with margin-aware mutual information regularization," *Proc Interspeech*, vol. 1641–5, 2022.