Ingénierie des Systèmes d'Information
Vol. 31, No. 8, August, 2026, pp. 2807-2816
Journal homepage: http://iieta.org/journals/isi

Evolution of Speaker Recognition: From Handcrafted Signal Processing to Deep Learning-Based Representation Learning

Hajer Y. Khdier Abbas Alkhalidy

Electronic Computer Center, Al Nahrain University, Baghdad 64074, Iraq

Corresponding Author Email: hajer91_ali@nahrainuniv.edu.iq

https://doi.org/10.18280/isi.310824

Received: 25 January 2026
Revised: 15 July 2026
Accepted: 20 August 2026
Available online: 31 August 2026

Keywords:
speaker recognition, deep learning, speaker embedding, x-vector, anti-spoofing, explainable artificial intelligence

ABSTRACT

Speaker recognition has shifted from handcrafted signal-processing techniques to powerful deep learning-based representation-learning methods. The paper provides a thorough review of speaker recognition systems' technological development, moving from traditional acoustic feature engineering to data-driven neural approaches. First, the classical Mel-Frequency Cepstral Coefficients (MFCCs) framework, Gaussian Mixture Model–Universal Background Model (GMM-UBM) framework, and the identity vector (i-vector)-Probabilistic Linear Discriminant Analysis (PLDA) framework are reviewed, and the principles and limitations of these approaches are discussed. In turn, the most recent deep learning architectures are briefly reviewed, such as Deep Neural Networks (DNNs), d-vector and x-vector models, end-to-end (E2E) learning, attention mechanisms, and Transformer models in speaker representation learning. In addition to the problem of recognition accuracy, there were other issues raised that are important for deployment in real-world scenarios, such as robustness to spoofing attempts, demographic fairness, the required computational efficiency, and the necessity for explainable artificial intelligence (XAI). Lastly, emerging research directions for self-supervised learning (SSL), multimodal biometric fusion, and lightweight deployment to edge devices are addressed. In structuring existing solutions in a single evolutionary perspective, researchers and practitioners are well armed with an understanding of the current state of speaker recognition technologies and the future opportunities they offer.

1. INTRODUCTION

Historically and commercially, voiceprint recognition is also called speaker recognition and is a foremost biometric modality which is based on the speaker's unique and recognizable characteristics for identification and/or authentication. In this review, the term speaker recognition will be used as the main technical term, equivalent to the term voiceprint, which is the biometric term for the same. It is a technology that is different from speech recognition – which aims to capture the spoken word – and particularly appealing for its non-intrusiveness, low cost, and ability to integrate into remote channels, including telephony and internet-based technology. That is the reason it is used in a variety of applications, such as access control in high-security areas, verification of financial transactions, forensic analysis, personalization of consumer electronics, and smart assistants. A remarkable technological evolution is revealed by tracing the field's growth from its humble beginnings in digital signal processing (DSP) to the present day, when artificial intelligence (AI) is built on deep learning architectures.

There are classical approaches that provided the groundwork for automated speaker recognition that involved a two-stage process known as feature extraction and back-end modeling. The first step in the field of feature extraction was the introduction of Mel-Frequency Cepstral Coefficients (MFCCs) [1]. MFCCs are able to give a compact and effective description of the power spectrum of the vocal tract in short durations by transforming the spectrum of the speech signal to the non-linear Mel scale, which mimics the frequency response of the human auditory system. The statistical approaches dominated the modeling phase for more than ten years; in this case, the distribution of the acoustic features of a speaker is modeled as a weighted sum of Gaussian densities, and verification is done by adapting a UBM (trained on a large population of speakers) to a claimant's speech and calculating a log-likelihood ratio. However, this approach, Gaussian Mixture Model-Universal Background Model (GMM-UBM), was very sensitive to session and channel variation [2]. To address this, researchers introduced Total Variability (TV) modeling, which led to the development of the "identity vector (i-vector)" [3]. This paradigm was the high watermark of the classical signal processing era, delivering a strong and compact representation that fuelled much of the progress in this area, and explicitly aiming to eliminate nuisance variability.

However, traditional methods were limited in terms of performance, especially in "in-the-wild" scenarios with high background noise, reverberation, cross-channel effects, and short-duration utterances, due to their dependence on

handcrafted features and statistical models that cannot sufficiently represent the high-dimensional, intricate, and non-linear nature of speaker-discriminative information. With the dawn of deep learning, driven by vast datasets and powerful computational resources, a paradigm shift was triggered. One of the most powerful capabilities of Deep Neural Networks (DNNs) was the data-driven and automatic learning of hierarchical and highly discriminative feature representations directly from raw or semi-processed speech signals [4], which meant that no manual feature engineering was required and more robust and abstract representations of the speaker's identity could be identified.

Deep learning was first utilized by employing DNNs as strong classifiers to generate embeddings of speakers. A first example of this was the "d-vector" system, in which the activations from a hidden layer of a DNN were extracted as a fixed-dimensional speaker representation [5], which was trained to classify speakers; subsequently, more elaborate architectures designed specifically for sequence data were developed. The combination of these properties resulted in the development of the "x-vector" architecture, a widely used contemporary baseline [6] due to the ability of Convolutional Neural Networks (CNNs) to capture local spectro-temporal patterns and of Recurrent Neural Networks (RNNs) and their variants, e.g., Long Short-Term Memory (LSTM) units, to model long-range temporal dependencies. X-vectors are generally created by forward-passing the features through the Time Delay Neural Network (TDNN) architecture and a statistical pooling layer, which consolidates frame-level features into a strong utterance-level embedding, especially built to support variable-length inputs.

There are two major trends in the current state of the art: end-to-end (E2E) systems and attention-based architectures. E2E models are trained to directly optimize the final verification metric (e.g., Equal Error Rate (EER)) and, as such, can reduce the pipeline and sometimes improve performance. At the same time, great success has been shown by the transformer architecture with its self-attention mechanism, as the model is able to dynamically weight the importance of each part of an utterance when generating the final speaker embedding. Furthermore, even more contextually rich [7] representations have been produced thanks to the abundance of large-scale, public datasets like VoxCeleb [8] that consist of millions of utterances from thousands of speakers.

Unlike previous surveys that focus primarily on isolated components of voiceprint recognition, a comprehensive paradigm-shifting roadmap is provided in this review. Traditional signal processing is contrasted with deep learning models under a unified taxonomical framework, highlighting the transition from handcrafted features to complex deep embeddings [9]. The theoretical principles and limitations of traditional signal processing methods, such as MFCCs and i-vectors, are first described, before the development of deep learning systems is systematically surveyed, starting from the simple d-vector models to the advanced x-vector and E2E architectures. Each stage is analyzed in comparison with one another to draw out the performance improvements and methodology changes. Finally, a structured review of recurring issues is provided, such as robustness to spoofing attacks, algorithmic fairness and bias, computational efficiency, and emerging future avenues like pre-trained foundation models that are developing the next generation of speaker (voiceprint) recognition technology.

2. CLASSICAL METHODOLOGIES IN SPEAKER RECOGNITION: A SIGNAL PROCESSING PERSPECTIVE

The classical methods were the foundations of the entire field of speaker recognition and are still of critical importance for understanding the field; therefore, these classical methods are reviewed in detail in this chapter, before the era of deep learning. The classical pipeline consisted of three steps: short-term acoustic feature extraction from the raw audio waveform, the construction of a mathematical model representing the distributional properties of the features of the speaker of interest, and the computation of a score that represents the likelihood of a match between a test utterance and a claimed identity.

2.1 Acoustic feature extraction: The foundation of voice representation

The first, and often most difficult, problem in any speaker recognition system in which structured models are used for recognition is the need to represent a raw one-dimensional audio waveform as a structure that retains the acoustic cues most useful for identifying the speaker. Furthermore, robustness is required from an ideal feature set against other sources of variation that may arise in the speech, such as the phonetic content thereof, which is irrelevant to the speaker's identity. MFCCs were considered the ultimate answer to this multifaceted problem that endured for many years, largely because of the elegant way that signal processing techniques and fundamental principles of human auditory perception were integrated.

The generation of MFCCs is a sequential process initiated by pre-emphasis, which is a first-order high-pass filter applied to the raw signal so that it is spectrally flattened and the higher frequencies are boosted. Next, the signal is divided into overlapping frames, of length 20–30 ms, during which the speech signal may be treated as quasi-stationary. Each frame is multiplied by a windowing function, such as the Hamming window, to taper its edges, thereby minimizing the spectral distortion that would otherwise arise from applying the Fourier transform to a non-periodic segment. The Fast Fourier Transform (FFT) is then applied to each windowed frame, transitioning the representation from the time domain to the frequency domain and yielding a magnitude spectrum. The core psychoacoustic innovation of MFCCs occurs in the next stage, where the linear frequency axis of the power spectrum is warped onto the non-linear Mel scale, which more accurately reflects the frequency resolution of the human cochlea, This warping is implemented by applying a bank of overlapping triangular filters to the power spectrum, producing a vector of log-energies for each Mel-spaced band, In the final step, the Discrete Cosine Transform (DCT) is applied to this vector of log-filterbank energies, The DCT is instrumental as a highly effective decorrelation of the energy vectors is performed, a property that is particularly beneficial for the diagonal covariance matrices commonly used in subsequent statistical models. To capture the dynamic evolution of the vocal tract, the first- and second-order temporal derivatives of these static coefficients, known as delta and double-delta features, are often computed and appended. Although other features like Linear Predictive Cepstral Coefficients (LPCCs) [10] and Perceptual Linear Prediction (PLP) coefficients [11] have also been developed,

MFCCs have consistently been regarded as the classical gold standard, offering the best performance/cost balance.

2.2 Early modeling techniques: Vector Quantization

After feature extraction, the first successful speaker modeling systems were based on Vector Quantization (VQ) [12]. Conceptually, VQ is a simple algorithm aimed at building a codebook from a speaker's enrollment speech. The process starts from an utterance set for a speaker, from which a series of MFCC vectors is extracted and then input to an iterative clustering algorithm. Most commonly, the Linde-Buzo-Gray (LBG) algorithm [13] is used, by which the high-dimensional feature space is partitioned into a predetermined number of clusters, and the centroid of each cluster is regarded as a "codeword" in the speaker's codebook. In verification, the test utterance is used to extract MFCCs. The minimum distortion (usually the squared Euclidean distance to the closest codeword) is then computed for each feature vector in this test sequence, and the final verification score is obtained as the average of all of these minimum distortions; if this average distortion is lower than a threshold, the speaker is accepted, otherwise rejected. Despite being simple and computationally efficient, a significant theoretical drawback is exhibited by the VQ method: each feature vector is assigned to a single codeword, without modeling the probabilistic nature of the classes and the inherent overlap found in the feature space. Hence, a search for more sophisticated, probabilistic modeling techniques was initiated,

2.3 Advanced statistical modeling: from Gaussian Mixture Models to i-vectors with linear discriminant backends

The way was thus paved for the use of more sophisticated probabilistic modeling techniques that were able to approximate the underlying probability density function (PDF) of the acoustic features of the speakers. The golden era of classical speaker recognition was marked by this transition, during which the GMM-UBM paradigm became the dominant approach [14]. A far more nuanced representation than VQ is provided by a GMM by modeling the feature distribution as a weighted sum of multiple multivariate Gaussian components; through this "soft" probabilistic assignment of feature vectors to mixture components, the complex, overlapping nature of a speaker's acoustic space is represented with much greater fidelity.

The true innovation of the framework, however, lay in its training and adaptation methodology. Instead of training a speaker model from a blank slate, a highly robust, speaker-independent GMM, known as the Universal Background Model (UBM), is first trained on a massive corpus of speech from a diverse population [8]. This UBM serves as a statistical representation of the general "anti-speaker." A specific speaker's model is then created not by de novo training, but through the adaptation of the UBM's parameters using the speaker's enrollment data, This is accomplished via Maximum A Posteriori (MAP) adaptation [15], a Bayesian technique that provides a principled way to update the UBM's parameters, Thus, the creation of well-estimated models is enabled even from short enrollment sessions, The final verification score is calculated as a Log-Likelihood Ratio (LLR), whereby the score is normalized by comparing the likelihood of the test utterance under the claimant's model against its likelihood under the general population model (UBM).

Despite its power, difficulty was encountered within the GMM-UBM framework in robustly disentangling the desired inter-speaker variability from the confounding intra-speaker or "session" variability [16]. This challenge was elegantly and powerfully addressed by the introduction of the Total Variability (TV) space and its product, the i-vector [17]. The core concepts, feature inputs, strengths, weaknesses, and references of the main classical speaker recognition models (ranging from VQ and GMM-UBM to the i-vector/Probabilistic Linear Discriminant Analysis (PLDA) framework) are summarized in Table 1.

Table 1. Comparison of advanced statistical speaker recognition models (GMM-UBM to i-vector/PLDA)

Methodology	Core Concept	Feature Input	Strengths	Weaknesses	Reference
Vector Quantization (VQ)	Speaker represented by a codebook of feature vector centroids	MFCCs	Simple to implement, computationally inexpensive	Poor statistical representation, sensitive to variations, low accuracy	N/A
GMM-UBM	Probabilistic density model of features, adapted from a universal background model	MFCCs	Strong statistical foundation, effective model adaptation (MAP), effective LLR scoring	Fails to disentangle speaker and channel effects, requires significant adaptation data	[16]
i-vector/PLDA	Utterance represented by a low-dimensional vector in a "TV" space	MFCCs	Fixed-length representation for variable-length audio, state-of-the-art for its era	Channel effects are mixed with speaker identity; performance degrades in short utterances	[17-19]

Note: GMM-UBM = Gaussian Mixture Model-Universal Background Model, PLDA = Probabilistic Linear Discriminant Analysis, MAP = Maximum A Posteriori, LLR = Log-Likelihood Ratio, MFCC = Mel-Frequency Cepstral Coefficients, TV = Total Variability.

The comparative analysis in Table 1 illustrates the clear evolutionary path of classical modeling. The transition from VQ to GMM-UBM introduced a probabilistic framework capable of modeling soft overlaps in acoustic space. However, both failed to decouple speaker characteristics from environmental and channel factors. This fundamental issue was resolved by the i-vector/PLDA framework, which mapped the acoustic session to a low-dimensional TV space, though at the expense of performance degradation on short-duration test utterances.

This approach shifted the modeling focus from the acoustic feature space to the GMM parameter space itself; the core assumption is that a speaker- and session-dependent GMM supervector, M, can be described as a linear deviation from the UBM's supervector, m, where this deviation lies within a single, low-dimensional subspace defined by a rectangular matrix T as in Eq. (1).

$$M = m + Tw \tag{1}$$

where, the vector w is the i-vector, a low-dimensional, fixed-length latent variable estimated for each utterance (Eq. (1)). However, the raw i-vector inherently conflates speaker characteristics with session effects. To isolate the speaker-specific information, a crucial final stage of channel compensation, or "backend" modeling, is required. The most successful linear model for this task is PLDA [18, 19]. PLDA is a generative model that provides a formal statistical framework for decomposing an i-vector into its constituent parts:

$$w = \mu + \Phi y + \epsilon \quad (2)$$

This is a fundamental linear model as in Eq. (2), where w is the observed i-vector, μ is the global mean, Φ is a matrix of columns that are the basis for the speaker variability subspace ("eigenvoices"), y is a latent variable containing the speaker's unique coordinates, and ϵ is the residual term corresponding to the session variability. A large set of development i-vectors is used to learn the parameters of the PLDA model. The task for verification is to compare two i-vectors by computing the log-likelihood ratio score, using the same-speaker and different-speaker hypotheses. This scoring function is useful to reduce such channel variability and emphasize the speaker variability. The i-vector/PLDA pipeline was therefore the culmination of the classical era [20] and was consistently shown to perform well in large-scale evaluations [21].

2.4 Limitations of classical approaches and the bridge to deep learning

The i-vector/PLDA system was the result of the classical pipeline over the years, which had a high level of statistical modeling. These approaches are based on assumptions that resulted in intrinsic limitations, especially in the case of “in the wild” data sets, which are unstructured. The most crucial of these assumptions was the reliance on handcrafted features such as MFCCs, which proved effective but were an information bottleneck. In addition, the statistical models were “shallow” in nature, lacking the ability to model the high non-linearity and hierarchical nature of voice production; the basis of distinguishing between the speaker's voiceprint and the extrinsic session variability was addressed by making linear approximations that were found to be unsuitable for real-world conditions. In combination with sub-optimal fixed features, shallow models' poor performance and the ongoing nuisance variability problem set a clear ceiling on system performance, and a new paradigm, the deep learning revolution [22], was introduced. However, the classical methods are still held in high regard today. Deep learning architectures are prone to severe overfitting in cases of very small datasets, while statistical models such as GMM-UBM can be used in a robust way, with no need for massive training corpora. Moreover, classical methods do not require heavy GPU resources and have a much lower memory footprint, making them perfect for lightweight deployment on edge devices and embedded systems, which are often resource-constrained. Last, they have clear mathematical underpinnings, allowing them to be used in fields like forensic science and legal audit, where ‘black-box’ deep learning decisions are not admissible as evidence.

Although the classical i-vector/PLDA pipeline was a spectacular success in statistical engineering, the pipeline itself was ultimately limited by the performance of its individual components—the handcrafted features and linear channel compensation. This inevitably limited performance in acoustically diverse and "in-the-wild" scenarios [21]. With the introduction of deep learning, spurred by the availability of vast amounts of data and incredible computational power provided by Graphics Processing Units (GPUs) did not just tweak the existing methods, but it also brought a paradigm shift to the field. This revolution is carefully mapped in this section, starting from the first use of DNNs, up to the creation of advanced, E2E architectures that now represent the state-of-the-art.

In summary, a rigid multi-stage pipeline relying on handcrafted feature extraction (MFCCs) and statistical modeling (VQ, GMM-UBM, and i-vector/PLDA) was established by classical methodologies. While mathematically elegant and computationally lightweight, these linear, shallow frameworks suffered from an information bottleneck and struggled to isolate speaker identity robustly from session and channel noise, thereby paving the way for data-driven deep learning representations.

3. DEEP LEARNING APPROACHES FOR SPEAKER RECOGNITION

A paradigm shift has occurred in the speaker recognition field through the use of deep learning techniques, whereby the capacity to obtain highly discriminative speaker representations has been greatly increased. Unlike handcrafted features like MFCCs and statistical modeling, the automatic learning of features directly from raw signals or preprocessed audio signals is enabled by deep learning methods. In this section, the development of deep-learning-based speaker recognition systems is discussed, starting from early neural network-based models and the introduction of d-vectors (Section 3.2), followed by the use of explicit models of temporal dynamics such as Convolutional and Recurrent architectures (Section 3.3).

The current baseline x-vector architecture is then discussed (Section 3.4), followed by a description of the design of loss functions that favor highly discriminative embeddings (Section 3.5). An overview of recent developments in speaker representation learning and the foundations of state-of-the-art systems is provided in this review, which are exhaustively covered by these subsections.

3.1 A paradigm shift from handcrafted features to learned representations

The most paradigm-shifting property of deep learning is its ability to learn representations. Optimal and hierarchical representations of features can be learned by DNNs directly from speech data, bypassing the information bottleneck of manual feature engineering, while complex, non-linear, and highly discriminative patterns in the speech signal (which may not be apparent to traditional algorithms [23]) are extracted. Initially, these networks were fed low-level acoustic features such as log Mel-filterbank energies, which are less heavily processed than MFCCs and thus retain more raw information, The deep, multi-layered structure of a neural network allows it to construct a hierarchy of features: simple, local spectro-temporal events (like edges or basic phonetic shapes) may be detected by the initial layers; these simple features are then composed by subsequent layers into more complex and abstract representations (such as formant transitions or

syllable-level patterns); and these abstract features are finally mapped into a space that is highly discriminative of speaker identity. This form of automated, data-driven feature learning, optimised concurrently with the classification task, is what makes deep learning-based systems perform so well and much more robust to noise and other acoustic variations, particularly the ability of the data sets to be trained without overfitting, through the use of large-scale, publicly available speech corpora and crucial techniques such as data augmentation [24, 25] (whereby the training set is artificially expanded by adding noise, reverberation or speed/pitch perturbations, significantly enhancing the model's ability to generalise).

3.2 Early architectures: Deep Neural Networks and the genesis of the d-vector

The first successful application of deep learning to the problem of speaker recognition was to repurpose architectures that have already achieved revolutionary results in the closely related problem of automatic speech recognition (ASR) [26]; the first approach was based on using a standard feed-forward DNN trained as a speaker classifier. The input to the network is a feature vector, consisting of features—such as 40-dimensional filterbanks—for a small temporal context window of neighboring frames, fed through multiple hidden layers that consist of non-linear activation functions such as Rectified Linear Unit (ReLU), and then through a Softmax output layer that calculates a probability distribution over a fixed number of speakers that were presented during the network's training.

The crucial innovation was not the classification output itself, but the realization that the activations of one of the final hidden layers could serve as a powerful, fixed-dimensional speaker-discriminative embedding [27]; this embedding was termed the d-vector, The rationale is that in order to correctly classify speakers at the output layer, the network is forced to learn an internal representation within its hidden layers where the feature vectors of different speakers are linearly separable, Therefore, the vector of activations from a bottleneck layer just before the final classifier encapsulates a rich, discriminative summary of the speaker's identity within that short time window. In the variant of this approach used for variable-length test utterances, d-vectors were extracted for each context window in turn, and the utterance-level representation was simply obtained by averaging these vectors over the entire length of the test sentence. The use of a fixed context window limited the network's capacity for modeling long-range temporal dependencies, and the simple averaging of the d-vectors across the entire utterance was a suboptimal pooling strategy, which treated all parts of the utterance as being of equal importance.

3.3 Modeling temporal dynamics: Convolutional and Recurrent architectures

To circumvent the deficiencies of standard DNNs, architectures naturally suited for sequential and structured data were adopted. Speech spectrograms were a natural fit for processing by CNNs, which had been used to reach superhuman performance in image recognition [28, 29]. A 2D spectrogram can be thought of as an image with one axis being the time axis and the other axis being the frequency axis. Filters (kernels) were learned by CNNs over this 2D representation, which allowed them to learn local spectro-temporal characteristics (such as formant trajectories, harmonic structures, and phonetic transitions), making their ability to learn localized, shift-invariant features a more powerful front-end than the fully-connected layers of a standard DNN.

At the same time, RNNs and particularly their gated counterparts, such as the LSTM and the Gated Recurrent Unit (GRU), were investigated as models of long-range temporal dependencies [30]. RNNs have a hidden state which functions as a memory, whereas feed-forward networks do not, making it possible to hold and accumulate information for an entire arbitrary-length utterance. Sophisticated designs are exhibited by the LSTM's input gate, forget gate, and output gate, enabling the model to determine what information to store, forget, or use to generate the current output, thus addressing the vanishing gradient problem that plagued simple RNNs [31]. In the typical LSTM-based speaker recognition system, the final hidden state of the LSTM after processing all the frames of an utterance is taken as the utterance-level speaker embedding, which means that theoretically the final state has captured a summary of the entire utterance. Furthermore, these models were improved by the addition of attention mechanisms, whereby weights are dynamically assigned to different frames during utterance embedding, thereby focusing on the speaker-discriminative part of the speech [32].

3.4 The modern baseline: The x-vector architecture

The x-vector system, typically integrated into the Kaldi ASR toolkit [33], is a carefully designed pipeline that integrates the best of previous systems and is currently the most influential and widely used deep learning-based speaker recognition system in the field today [34], comprising three key components.

The first is a feature extractor, which is usually realized as a TDNN, operating at the frame level. A TDNN is effectively a one-dimensional Convolutional Neural Network (1D-CNN) applied over the temporal axis, where each layer processes a spliced context of frames from the layer below. This structure is computationally efficient and highly effective at learning temporal patterns in speech; the initial layers of the TDNN extract low-level acoustic features, while deeper layers learn more complex and context-aware representations.

The second and most innovative component is the statistics pooling layer. After the frame-level TDNN processing, instead of simple averaging (as in d-vectors) or taking the last state (as in some RNN approaches), the frame-level representations are aggregated over the entire utterance by this layer through computing their first- and second-order statistics: the mean and the standard deviation. The motivation is that the mean vector captures the average acoustic characteristics of the speaker, while the standard deviation vector captures the variability or "spread" of their phonetic realizations, which is itself a powerful speaker-discriminative cue.

The third component is an utterance-level classifier. The concatenated mean and standard deviation vectors are passed through several additional feed-forward layers to produce the final fixed-dimensional embedding, which is the x-vector. The entire network is trained on a multi-class speaker classification task; the extracted x-vector is then typically processed by a backend scoring model, such as PLDA, which has proven to be remarkably effective even when applied to these deep-learned embeddings, whereby system robustness is further enhanced [35, 36].

3.5 Loss functions for highly discriminative embeddings

The choice of loss function used to train the embedding extractor is of paramount importance, as it directly shapes the geometry of the embedding space; the standard approach is to use the Softmax function coupled with cross-entropy loss. For a given training sample x_i with label y_i, the loss is:

$$L_{Softmax} = -\left(\frac{1}{N}\right)\sum_{i=1}^{N}\log\left[\frac{\exp\left(W_{\{y_i\}}^{T}x_i + b_{\{y_i\}}\right)}{\sum_{j=1}^{C}\exp\left(W_j^{T}x_i + b_j\right)}\right] \quad (3)$$

where, x_i is the deep embedding, and W_j and b_j are the weights and bias for the j-th class in the final classification layer as in Eq. (3). While this loss function effectively enforces separability (i.e., it pushes embeddings from different classes apart), it does not explicitly encourage two crucial properties for open-set speaker verification: intra-class compactness (forcing embeddings from the same speaker to be tightly clustered) and inter-class margin (enforcing a larger separation between different speaker clusters).

To address this, a new family of margin-based Softmax losses was developed, which has become the state-of-the-art [37]. These losses work by reformulating the logit $W_j^T x_i$ in terms of cosine similarity, $||W_j||\ ||x_i||\ cos(\theta_j)$, where θ_j is the angle between the embedding x_i and the weight vector for class j. By fixing $||W_j|| = 1$ and normalizing the embedding $||x_i|| = s$. Consequently, the optimization process shifts toward regulating angular distances on a unit hypersphere. A prominent loss function employing this formulation is the Additive Angular Margin Softmax (AAM-Softmax), widely known as ArcFace [38], which introduces an explicit additive margin in the angular space to enforce tight intra-class compactness and maximize inter-class separability [38]. It introduces an AAM, m, directly into the angle of the target class as in Eq. (4):

$$L_{AAM} = -\left(\frac{1}{N}\right)\sum_{i=1}^{N}\log\left[\frac{\exp\left(s\cos\left(\theta_{\{y_i\}}+m\right)\right)}{\exp\left(s\cos\left(\theta_{\{y_i\}}+m\right)\right)+\sum_{j=1,j\neq y_i}^{C}\exp\left(s\cos(\theta_j)\right)}\right] \quad (4)$$

This modification makes the optimization problem more difficult, forcing the network to learn features that are not only separable but also highly compact and widely separated in the angular domain; this results in embeddings with significantly improved discriminative power. Other popular variants include CosFace [39], which adds a margin to the cosine value, and SphereFace [40], by applying a multiplicative angular margin. Ultimately, adopting these sophisticated margin-based loss functions has been instrumental in driving state-of-the-art performance in deep speaker recognition architectures [41-44].

As summarized in Table 2, the development of DNN architectures addresses different trade-offs between computational complexity and temporal modeling. While d-vectors paved the way, their reliance on simple frame averaging lost critical dynamic sequence details. CNNs and RNNs introduced localized and sequential modeling, respectively, but were surpassed by TDNN-based x-vectors. The x-vector's integration of a statistics pooling layer (mean and standard deviation) successfully captures both the distribution and variance of speaker characteristics across variable-length utterances.

Table 2. Comparison of deep learning architectures and pooling methods for speaker embedding extraction

Methodology	Core Architecture	Pooling/Aggregation	Key Strengths	Primary Weaknesses
DNN/d-vector	Feed-forward DNN	Temporal averaging of frame embeddings	Pioneered deep embeddings; simple to implement	Limited temporal context; suboptimal pooling
CNN-based	CNNs (1D or 2D)	Temporal averaging or global pooling	Learns local, shift-invariant spectro-temporal patterns	Can struggle with very long-range dependencies
RNN/LSTM-based	LSTM or GRU	Final hidden state or attention pooling	Explicitly models long-range temporal sequences	Computationally more intensive; can focus too much on the last frames
TDNN/x-vector	TDNN	Statistics pooling (mean + standard deviation)	Highly effective pipeline; pooling captures feature distribution	Still benefits from a separate backend (PLDA); complex pipeline
Margin-based methods	Any deep architecture (e.g., TDNN, ResNet)	Statistics pooling or attention	Loss function (e.g., ArcFace) creates a highly discriminative embedding space	Requires large, clean training sets for optimal margin learning

Note: DNN = Deep Neural Network; CNN = Convolutional Neural Network; RNN = Recurrent Neural Network; LSTM = Long Short-Term Memory; TDNN = Time Delay Neural Network; GRU = Gated Recurrent Unit; PLDA = Probabilistic Linear Discriminant Analysis.

Table 2 shows an overview of the deep learning architectures that have been used for speaker embedding extraction, the pooling steps used, their key strengths, and their primary weaknesses. This overview places the design decisions made in the highly discriminative embeddings and the subsequent use of margin-based loss functions for open-set speaker verification in context.

In summary, deep learning resolved the limitations of handcrafted feature engineering by enabling automatic, hierarchical representation learning. The evolution from basic d-vectors to statistical-pooling-based x-vectors significantly improved temporal context modeling. When integrated with modern margin-based loss functions such as AAM-Softmax, these models compress intra-class variability while expanding inter-class separation, establishing highly robust baselines for open-set speaker verification.

4. MODERN TRENDS AND ADVANCED ARCHITECTURES

The x-vector pipeline, which was supported by margin-based loss functions, was a major advancement in deep learning for speaker recognition. But the research community kept identifying and dealing with the residual architectural suboptimalities, and more recently, the prevailing trends of simplifying the recognition pipeline and strengthening the model's capacity to interpret contextual information over an

entire utterance have resulted in the creation of fully E2E systems and the infiltration of powerful attention-based mechanisms, such as the Transformer architecture.

4.1 From multi-stage pipelines to end-to-end systems

A critical observation of the x-vector/PLDA pipeline is its multi-stage, disjointed nature: the DNN is trained as a speaker classifier (a proxy task), while the backend PLDA model is trained separately to perform the actual verification scoring. The front-end embedding extractor is therefore not directly optimized for the final evaluation metric, such as the EER or the minimum Detection Cost Function (minDCF); this disconnect is inherently suboptimal. E2E speaker verification aims to resolve this by training a single neural network directly to optimize a verification-based loss function; this is typically achieved through metric learning, where the network learns a function that maps utterances to an embedding space where the distance between embeddings corresponds directly to similarity.

Two prominent loss functions have enabled this E2E paradigm; the first is Contrastive Loss, which operates on pairs of utterances. For a pair of embeddings (w_i, w_j) with a label Y in {0, 1} (where $Y = 1$ if they are from the same speaker and $Y = 0$ otherwise), the loss is:

$$L_{contrastive} = Y\left(\frac{1}{2}\right)(D_w)^2 + (1-Y)\left(\frac{1}{2}\right)\{\max(0, m - D_w)\}^2 \quad (5)$$

In this loss function, $D_w = \left|\left|w_i - w_j\right|\right|$ is the Euclidean distance, and m is a predefined margin. This loss function pulls same-speaker pairs together and pushes different-speaker pairs apart until they are separated by more than margin m in Eq. (5).

The Triplet Loss is a stronger formulation, based on a triplet of embeddings: an "anchor" (w_a), a "positive" from the same speaker (w_p), and a "negative" from a different speaker (w_n). The loss is defined as:

$$L_{triplet} = \sum_{i=1}^{N}\left[\left|\left|w_a^i - w_p^i\right|\right|^2 - \left|\left|w_a^i - w_n^i\right|\right|^2 + \alpha\right] \quad (6)$$

This term $[x]+$ translates to $\max(0, x)$, and the margin hyperparameter α is directly enforced by the loss, which corresponds to the desired geometric property: the distance between the anchor and the positive is enforced to be less than or equal to the distance between the anchor and the negative by at least the margin α in Eq. (6).

Training a system using these losses simplifies the deployment pipeline by avoiding the need for a separate PLDA backend, and can result in better embeddings, since everything, from feature extraction to scoring, is optimized for the same purpose: speaker discrimination.

4.2 The rise of attention and transformer-based models

One of the main problems with the statistics pooling layer in the x-vector system is that it simply computes an unweighted mean and standard deviation for each frame of an utterance. Clearly, not all frames are equally informative: frames of silence and background noise provide little speaker-discriminative information compared to frames of clear voiced speech or of unvoiced fricatives. One of the most powerful solutions to this problem is through attention mechanisms, which are data-driven approaches. This enables the model to learn to give a set of importance weights to the sequence of feature vectors extracted from each frame, and produce the utterance-level embedding as a weighted average, thus focusing on the most speaker-discriminative parts of speech and ignoring the less relevant parts.

Taking this concept to its logical conclusion has led to the adoption of the transformer architecture, which relies entirely on a mechanism known as self-attention. A transformer-based encoder processes the entire sequence of frame-level features simultaneously. Its self-attention layers allow each frame to compute its updated representation by attending to every other frame in the sequence; this enables the model to capture extremely complex, long-range contextual relationships and dependencies in a highly parallelizable manner, overcoming the sequential processing bottleneck of RNNs. For speaker recognition, this enables the model to capture, for example, a correlation between a specific formant structure within a part of an utterance and a specific intonation pattern at a later point in the same utterance, and to encode these global cues into a very context-aware and strong speaker embedding. Recently, the various types of transformers and their architectures have been widely used in many cutting-edge systems, where they have replaced feature extractors such as TDNN or RNN and achieved better performance, notably for variable-duration and noisy utterances.

5. PERSISTENT CHALLENGES AND FUTURE RESEARCH DIRECTIONS

While the strides made with deep learning are impressive, there are still some key issues that are at the forefront of speaker recognition research. Looking ahead, the field is increasingly concerned not only with accuracy, but also with the security, fairness, and transparency of these powerful biometric systems.

5.1 System robustness and anti-spoofing

Deep models work well when it comes to massive amounts of in-domain data, but are less effective when data is mismatched. There are additional issues that are still under investigation to improve robustness to real-world variability, like short-duration utterances, far-field audio with heavy reverberation, and unseen device channels. But a far more troubling and insidious problem is spoofing attacks – those that try to trick a biometric system with fake voice signals. These attacks are generally considered presentation attacks, and they can be classified as follows:

•Replay Attacks: Replaying a recording of a real user's voice.

•Text-to-Speech (TTS) Synthesis: Employing a synthesis engine that produces speech output that is similar to that of a target speaker's voice.

•Voice Conversion (VC): Alteration of the voice of a source speaker so as to sound like a target speaker.

The newness and advancement of such generative models create a significant security threat; thus, the concept of anti-spoofing, commonly referred to as presentation attack

detection (PAD), has been introduced into the scene. A countermeasure (CM) system is a binary classifier that can classify genuine (bona fide) speech versus spoofed speech. Initial methods considered the speaker verification module and the CM module as independent, and their scores were combined at the end. Such multi-task training can yield more secure and integrated systems; however, a tough race is currently developing between the continuous progress of speech synthesis and replay technology and the development of robust CMs.

5.2 Fairness, algorithmic bias, and explainability

Speaker Recognition systems are increasingly used in critical applications such as financial services and access control, leading to significant ethical concerns. One of the concerns is algorithmic bias. Because they rely on big data, deep learning models can learn and magnify biases in the data they receive. A system that is demographically imbalanced due to an uneven number of speakers in a training corpus (e.g., more male speakers than female speakers, or more speakers of one accent than another) will also be an unethical and dysfunctional one, because the system may have a high error rate on the underrepresented group (e.g., high error rate on female speakers, or high error rate on speakers of a different accent) and thus be unsuitable for that part of its user base. Research on this topic is primarily centered on curating more balanced datasets, creating data augmentation and re-weighting techniques to reduce bias, and designing optimization algorithms that are fair to the users.

A similar challenge is explainable artificial intelligence (XAI). Speaker recognition systems of today are indeed 'black boxes'. They generate a score, but they cannot explain why a certain 'decision' has been made; therefore, this lack of transparency is a big hurdle when it comes to user trust, system debugging, and regulatory compliance. Research in XAI for speaker recognition seeks to create methods for understanding the speaker recognition model's reasoning, including the creation of saliency maps that highlight those parts of the spectrogram that were most important for a decision to be made. One of the most critical areas for voice biometric systems in the future is ensuring that they are accurate, secure, fair, and transparent.

5.3 Emerging paradigms: Pre-trained speech models, multimodality, and edge deployment

Beyond the traditional deep learning pipelines, several modern paradigms are actively shaping the future of speaker recognition. First, self-supervised learning (SSL) via large pre-trained speech foundation models (such as wav2vec 2.0 and HuBERT) has emerged as a dominant force. These models leverage hundreds of thousands of hours of unlabeled audio to extract general speech representations, which can then be fine-tuned with minimal target speaker data. Second, multimodal biometric systems are gaining attention by fusing voiceprint recognition with facial and lip-movement features to guarantee security in highly noisy environments. Lastly, to support the deployment of state-of-the-art architectures on mobile and IoT devices, research into lightweight deployment techniques—such as knowledge distillation, network pruning, and float quantization—has become essential to optimize inference speed and energy efficiency without sacrificing performance.

6. CONCLUSION

This review started with the mathematically oriented and limited approach of classical signal processing and the use of MFCC features and statistical voiceprint models; it then covered the use of the very effective i-vector/PLDA pipeline; next, it traced the paradigm shift that began with the introduction of deep learning, replacing handcrafted features with learned representations. This eventually led to the development of increasingly complex deep learning architectures that ranged from early DNN-based d-vectors to the TDNN-based x-vector system that is now a modern industry baseline.

These models, which are trained using high-quality, margin-based loss functions on gigantic amounts of data, have reached unprecedented levels of accuracy, moving towards more architectural elegance and contextual awareness, such as the implementation of E2E, metric-learning-based systems, attention mechanisms, and transformer-based encoders. However, as this review has demonstrated, the future of voice biometrics is driven by the need to improve the capabilities of our models to become even more powerful, but also by the need to ensure that our models are more secure and fair against more sophisticated attacks, and that our models are explainable, transparent, and trustworthy for all users.

REFERENCES

[1] Janke, M., Diener, L. (2017). EMG-to-speech: Direct generation of speech from facial electromyographic signals. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(12): 2375-2385. https://doi.org/10.1109/TASLP.2017.2738568

[2] Jong, N.S., Phukpattaranont, P. (2019). A speech recognition system based on electromyography for the rehabilitation of dysarthric patients: A Thai syllable study. Biocybernetics and Biomedical Engineering, 39(1): 234-245. https://doi.org/10.1016/j.bbe.2018.11.010

[3] Lee, W., Seong, J.J., Ozlu, B., Shim, B.S., Marakhimov, A., Lee, S. (2021). Biosignal sensors and deep learning-based speech recognition: A review. Sensors, 21(4): 1399. https://doi.org/10.3390/s21041399

[4] Gaddy, D., Klein, D. (2020). Digital voicing of silent speech. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 5521-5530. https://doi.org/10.18653/v1/2020.emnlp-main.445

[5] Fiedler, L., Wöstmann, M., Graversen, C., Brandmeyer, A., Lunner, T., Obleser, J. (2017). Single-channel in-ear-EEG detects the focus of auditory attention to concurrent tone streams and mixed speech. Journal of Neural Engineering, 14(3): 036020. https://doi.org/10.1088/1741-2552/aa66dd

[6] Pinheiro, A.P., Schwartze, M., Kotz, S.A. (2018). Voice-selective prediction alterations in nonclinical voice hearers. Scientific Reports, 8: 14717. https://doi.org/10.1038/s41598-018-32614-9

[7] Sebkhi, N., Yunusova, Y., Ghovanloo, M. (2018). Towards phoneme landmarks identification for American-English using a multimodal speech capture system. In 2018 IEEE Biomedical Circuits and Systems Conference (BioCAS), Cleveland, USA, pp. 1-4.

https://doi.org/10.1109/BIOCAS.2018.8584737
[8] Manoni, L., Turchetti, C., Falaschetti, L., Crippa, P. (2019). A comparative study of computational methods for compressed sensing reconstruction of EMG signal. Sensors, 19(16): 3531. https://doi.org/10.3390/s19163531
[9] Poncela, A., Gallardo-Estrella, L. (2015). Command-based voice teleoperation of a mobile robot via a human-robot interface. Robotica, 33(1): 1-18. https://doi.org/10.1017/S0263574714000010
[10] Hwang, S., Jin, Y.G., Shin, J.W. (2019). Dual microphone voice activity detection based on reliable spatial cues. Sensors, 19(14): 3056. https://doi.org/10.3390/s19143056
[11] Maas, A.L., Qi, P., Xie, Z., et al. (2017). Building DNN acoustic models for large vocabulary speech recognition. Computer Speech & Language, 41: 195-213. https://doi.org/10.1016/j.csl.2016.06.007
[12] Ravanelli, M., Omologo, M. (2017). Contaminated speech training methods for robust DNN-HMM distant speech recognition. arXiv preprint arXiv:1710.03538. https://doi.org/10.48550/arXiv.1710.03538
[13] Zeyer, A., Irie, K., Schlüter, R., Ney, H. (2018). Improved training of end-to-end attention models for speech recognition. In Interspeech 2018, Hyderabad, India, pp. 7-11. https://doi.org/10.21437/Interspeech.2018-1616
[14] Hori, T., Cho, J., Watanabe, S. (2018). End-to-end speech recognition with word-based RNN language models. In 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, pp. 389-396. https://doi.org/10.1109/SLT.2018.8639693
[15] Sak, H., Senior, A., Rao, K., Beaufays, F. (2015). Fast and accurate recurrent neural network acoustic models for speech recognition. arXiv preprint arXiv:1507.06947. https://doi.org/10.48550/arXiv.1507.06947
[16] Takahashi, N., Gygli, M., Van Gool, L. (2018). AENet: Learning deep audio features for video analysis. IEEE Transactions on Multimedia, 20(3): 513-524. https://doi.org/10.1109/TMM.2017.2751969
[17] Amodei, D., Ananthanarayanan, S., Anubhai, R., et al. (2016). Deep Speech 2: End-to-end speech recognition in English and Mandarin. In Proceedings of the 33rd International Conference on Machine Learning, pp. 173-182. https://proceedings.mlr.press/v48/amodei16.html.
[18] Assael, Y.M., Shillingford, B., Whiteson, S., de Freitas, N. (2016). LipNet: End-to-end sentence-level lipreading. arXiv preprint arXiv:1611.01599. https://doi.org/10.48550/arXiv.1611.01599
[19] Ephrat, A., Peleg, S. (2017). Vid2speech: speech reconstruction from silent video. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, pp. 5095-5099. https://doi.org/10.1109/ICASSP.2017.7953127
[20] Biadsy, F., Weiss, R.J., Moreno, P.J., Kanvesky, D., Jia, Y. (2019). Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation. In Interspeech 2019, Graz, Austria, pp. 4115-4119. https://doi.org/10.21437/Interspeech.2019-1789
[21] Sun, C.W., Yang, Y.X., Wen, C., Xie, K., Wen, F.Q. (2018). Voiceprint identification for limited dataset using the deep migration hybrid model based on transfer learning. Sensors, 18(7): 2399. https://doi.org/10.3390/s18072399
[22] Chen, Y.C., Yang, Z., Yeh, C.F., Jain, M., Seltzer, M.L. (2020). Aipnet: Generative adversarial pre-training of accent-invariant networks for end-to-end speech recognition. In 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, pp. 6979-6983. https://doi.org/10.1109/ICASSP40776.2020.9053098
[23] de Almeida, F.L., Rosa, R.L., Rodriguez, D.Z. (2018). Voice quality assessment in communication services using deep learning. In 2018 15th International Symposium on Wireless Communication Systems (ISWCS), Lisbon, Portugal, pp. 1-6. https://doi.org/10.1109/ISWCS.2018.8491055
[24] Panayotov, V., Chen, G.G., Povey, D., Khudanpur, S. (2015). LibriSpeech: An ASR corpus based on public domain audio books. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, pp. 5206-5210. https://doi.org/10.1109/ICASSP.2015.7178964
[25] Lu, Y.Y., Li, H.B. (2019). Automatic lip-reading system based on deep convolutional neural network and attention-based long short-term memory. Applied Sciences, 9(8): 1599. https://doi.org/10.3390/app9081599
[26] Gosztolya, G., Pintér, Á., Tóth, L., Grósz, T., Markó, A., Csapó, T.G. (2019). Autoencoder-based articulatory-to-acoustic mapping for ultrasound silent speech interfaces. In 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, pp. 1-8. https://doi.org/10.1109/IJCNN.2019.8852153
[27] Algabri, M., Mathkour, H., Alsulaiman, M.M., Bencherif, M.A. (2021). Deep learning-based detection of articulatory features in Arabic and English speech. Sensors, 21(4): 1205. https://doi.org/10.3390/s21041205
[28] Akbari, H., Arora, H., Cao, L., Mesgarani, N. (2018). Lip2AudSpec: Speech reconstruction from silent lip movements video. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, Canada, pp. 2516-2520. https://doi.org/10.1109/ICASSP.2018.8461856
[29] Fernández-López, A., Sukno, F.M. (2018). Survey on automatic lip-reading in the era of deep learning. Image and Vision Computing, 78: 53-72. https://doi.org/10.1016/j.imavis.2018.07.002
[30] Hao, M., Mamut, M., Yadikar, N., Aysa, A., Ubul, K. (2020). A survey of research on lipreading technology. IEEE Access, 8: 204518-204544. https://doi.org/10.1109/ACCESS.2020.3036865
[31] Fernandez-Lopez, A., Martinez, O., Sukno, F.M. (2017). Towards estimating the upper bound of visual-speech recognition: The visual lip-reading feasibility database. In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, USA, pp. 208-215. https://doi.org/10.1109/FG.2017.34
[32] Eom, C.S.H., Lee, C.C., Lee, W., Leung, C.K. (2019). Effective privacy-preserving data publishing by vectorization. Information Sciences, 527: 311-328. https://doi.org/10.1016/j.ins.2019.09.035
[33] González-López, J.A., Gómez-Alanis, A., Martín Doñas, J.M., Pérez-Córdoba, J.L., Gómez, A.M. (2020). Silent

speech interfaces for speech restoration: A review. IEEE Access, 8: 177995-178021. https://doi.org/10.1109/ACCESS.2020.3026579
[34] Wang, J., Hahm, S. (2015). Speaker-independent silent speech recognition with across-speaker articulatory normalization and speaker adaptive training. In Proceedings of Interspeech 2015, pp. 2415-2419. https://doi.org/10.21437/Interspeech.2015-522
[35] Kapur, A., Kapur, S., Maes, P. (2018). AlterEgo: A personalized wearable silent speech interface. In Proceedings of the 23rd International Conference on Intelligent User Interfaces, Tokyo, Japan, pp. 43-53. https://doi.org/10.1145/3172944.3172977
[36] Kimura, N., Kono, M., Rekimoto, J. (2019). Sottovoce: An ultrasound imaging-based silent speech interaction using deep neural networks. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, Scotland, UK, pp. 1-11. https://doi.org/10.1145/3290605.3300376
[37] Bi, W., Zhang, C., Fu, G., Wang, M., Guo, Z. (2026). High-precision permanent magnet localization using an improved artificial lemming algorithm integrated with Levenberg–Marquardt Optimization. Electronics, 15(1): 135. https://doi.org/10.3390/electronics15010135
[38] Ibrahimov, I., Gosztolya, G., Csapó, T.G. (2024). Towards cross-speaker articulation-to-speech synthesis using dynamic time warping alignment on speech signals. In 2nd Workshop on Intelligent Infocommunication Networks, Systems and Services, pp. 7-12. https://doi.org/10.3311/wins2024-002
[39] Csapó, T.G., Al-Radhi, M.S., Németh, G., et al. (2019). Ultrasound-based silent speech interface built on a continuous vocoder. arXiv preprint arXiv:1906.09885. https://doi.org/10.48550/arXiv.1906.09885
[40] Salomons, I., del Blanco, E., Navas, E., Hernáez, I. (2025). Electrode setup for electromyography-based silent speech interfaces: A pilot study. Sensors, 25(3): 781. https://doi.org/10.3390/s25030781
[41] van den Oord, A., Dieleman, S., Zen, H., et al. (2016). WaveNet: A generative model for raw audio. In Proceedings of 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9), p. 125. https://www.isca-archive.org/ssw_2016/vandenoord16_ssw.html.
[42] Boles, A., Rad, P. (2017). Voice biometrics: Deep learning-based voiceprint authentication system. In 2017 12th System of Systems Engineering Conference (SoSE), Waikoloa, USA, pp. 1-6. https://doi.org/10.1109/SYSOSE.2017.7994971
[43] Choi, H., Jung, S., Sohn, J., Im, C.H. (2026). Efficient sensor configuration for accelerometer-based silent speech interfaces: Regionally distributed sensor placement strategy considering articulatory dynamics. Research Square Preprint. https://doi.org/10.21203/rs.3.rs-10442598/v1
[44] Kim, M. J., Cao, B.M., Mau, T., Wang, J. (2017). Multiview representation learning via deep CCA for silent speech recognition. In Proceedings of Interspeech 2017, pp. 2769-2773. https://doi.org/10.21437/Interspeech.2017-952

NOMENCLATURE

M	linear deviation from the UBM's supervector
w	i-vector
m	deviation lies within a single
T	low-dimensional subspace defined by a rectangular matrix
y	speaker's unique coordinates
L	Softmax function
x_i	deep embedding
W_j	weights
b_j	bias
DW	Euclidean distance
w_a	triplet of embeddings: an "anchor"
w_p	Form a same speaker
w_n	Form a different speaker
$[x]_+$	translates to max (x, 0)

Greek symbols

Φ	basis for the speaker variability subspace
ϵ	is the residual term corresponding to the session variability
θ_j	is the angle between the embedding x_i and the weight vector for class j
α	margin hyperparameter is directly enforced by the loss

Subscripts

j	j-th class in the final classification layer