An Overview of the Development of Speaker Recognition Techniques for Various Applications

Amira A. Mohamed[1,2], Amira Eltokhy[3], and Abdelhalim A. Zekry[1]
[1] Ain Shams University, Cairo 11517, Egypt
[2] Badr University in Cairo (BUC), Cairo 11829, Egypt
[3] MSA University, Cairo 12451, Egypt
Email: eng.amira.sakr@gmail.com; a.eltokhy@rapidbiolabs.com; aaazekry@hotmail.com; amira.ahmed@buc.edu.eg

Abstract—Speech Enhancement (SE) is a significant research issue in audio signal processing where the goal is to enhance the clarity and quality of speech signals corrupted by noise. Because of its different applications, it becomes a compelling research topic nowadays. The focus of this paper is dedicated to one of these applications which is Speaker recognition. In this paper, the fundamentals and applications of speaker recognition are discussed. A brief study on the performance and the recognition accuracy of different speaker modeling techniques has been conducted. Furthermore, while there have been several studies about the technologies used in speaker recognition, there have been few studies about the applications of speaker recognition, and none of them considers combining or linking the applications and technology in the same study. This overview demonstrates the various technologies that can be used to achieve speaker recognition applications. It aims to give a new perspective of the existing technologies uses in various applications. The paper concludes with discussions on future trends and research opportunities in this area.

Index Terms—Speaker recognition, speaker Modeling techniques, deep learning, speaker recognition applications.

I. Introduction

Speech is the most common and important way of communication among humans. Speech enhancement [1] is a critical requirement in the field of speech signal processing. Voice recognition technology can be broken down into two categories: speech recognition and speaker recognition. Speech recognition is a method of analysing the content of a speaker's words or speech [2], while speaker recognition, is the process of identifying people based on their voices. In this article, we are primarily focused on speaker recognition. Because of physical differences such as larynx sizes and vocal tract forms, no two people have the same sound. Furthermore, each speaker has a distinct voice and a manner of speaking, which includes the use of a specific accent, rhythm, intonation style, and pronunciation pattern. Most of these features are often used in speaker recognition systems, and they are used in a variety of ways to achieve more accurate recognition [3].

Based on the existence of speaker voice prints in a database, the recognition systems can be classified into closed-set recognition and open-set recognition. The closed-set refers to situations in which the unknown voice must come from a set of registered speakers, whereas the open-set refers to situations in which the unknown voice could come from unregistered speakers, in which case, this identification system could provide a "none of the above" option. Moreover, in practice, speaker recognition systems can be classified, based on the speech modalities or quality of speech, into text-dependent and text-independent recognition. Speakers in text-dependent Speaker Recognition Systems (SRS) are only able to say certain sentences or terms that the system recognises. These recognition phrases are pre-programmed or pre-determined. The text-independent SRS, on the other hand, could process freely spoken expression, which could be either a user-selected phrase or conversational speech. Text-independent SRS are more flexible, but also more complicated than text-dependent SRS [4]. It is regarded as the more difficult of the two tasks. Furthermore, text-independent systems are more commercially appealing than text-dependent systems in real life [5] because it is more difficult to imitate an unknown phrase than a known one [2].

As shown in the following sections, all the above speaker recognition classifications can be divided into two categories: speaker identification and speaker verification.

A. Speaker Identification

In speaker identification, human speech from an individual is used to identify who that individual is. (See Fig. 1). Training (also known as enrolment) is the process of acquiring speech from each known, verified speaker, for all speakers who need to be identified in order to build (train) the model for that speaker. This is usually done before the system is implemented and off-line as part of the system setup.

The true operation of the system is tested by comparing speech from an unknown utterance to each of the trained speaker models. In closed-set identification, the unknown person belongs to a pre-existing pool or database of speakers (speaker models), and the problem then becomes determining which speaker from the pool the unknown speech is extracted from.

The recognition rate is the most important performance indicator for such systems (percentage of correct identification averaged across all speakers in the pool). Closed-set identification is common in departmental organisations where community members are identified, speaker profiles may be obtained and retained in a database, and identification is limited to the department

Manuscript received January 5, 2022; revised July 15, 2022.
doi:10.12720/jcm.17.8.632-642

(i.e., there are no "external" users). The unknown person in open-set identification may come from the general population. However, because identification is always performed against a finite, known pool of individuals, arbitrary people cannot be identified. An open-set identification system's first task is to determine if the speaker belongs to a pool or database of known speakers; if not, the speaker is rejected; otherwise, closed-set identification is performed. It's critical in these systems to determine if a speaker is a member of the pool; otherwise, a random person from the pool will always be identified [6].

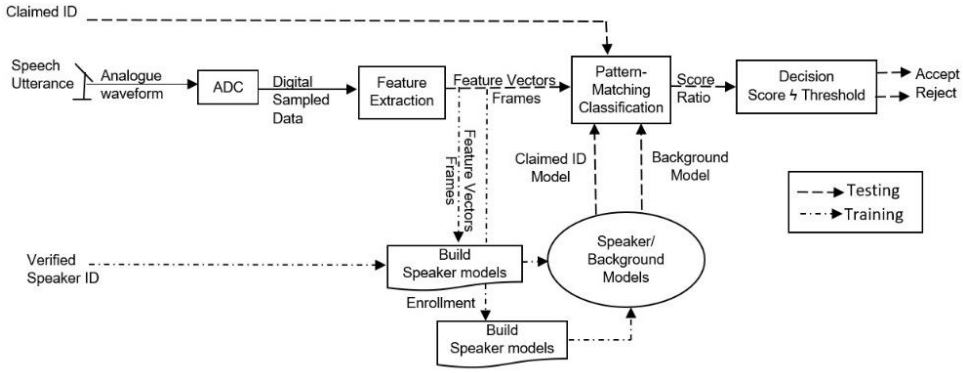

Fig. 1. Speaker identification system block diagram[6].

B. Speaker Verification

In speaker verification, human speech from an individual is used to verify the asserted identity of that individual (see Fig. 2). As with speaker identification, the system's initial configuration is carried out during training or enrolment, each speaker in order to be verified by the system must provide speech samples, which are then used to train the model for that speaker. In testing, verification occurs when the individual is required to make a claim about who he or she is, and the system then verifies whether that claim is true or false. With speaker verification, the unknown person's speech is compared to both the claimed identity and all other speakers (the imposter or background model(s)). The ratio of the two measures is then calculated and compared to a threshold; if it is greater than the threshold, the claim is accepted as true; if it is less than the threshold, the claim is rejected and is considered as false [6].

II. Feature Extraction

A. Feature Extraction Process

The extraction of vectors of features uniformly distributed over time from the time-domain sampled acoustic waveform is the most fundamental process shared by all types of speaker and speech recognition systems. Regardless of the features derived from the waveform (which are numerous), the initial framing of the waveform, as shown in Fig. 3, proceeds as follows in the coming three paragraphs (the numerical parameter values mentioned are those commonly used in practice)[6]:

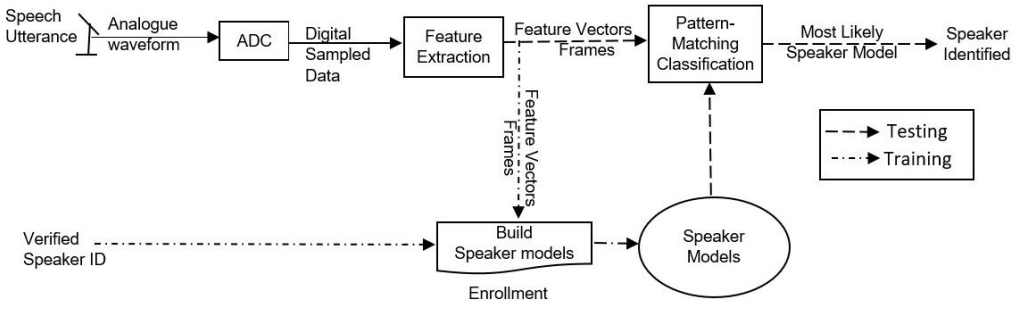

Fig. 2. Speaker verification system block diagram [6].

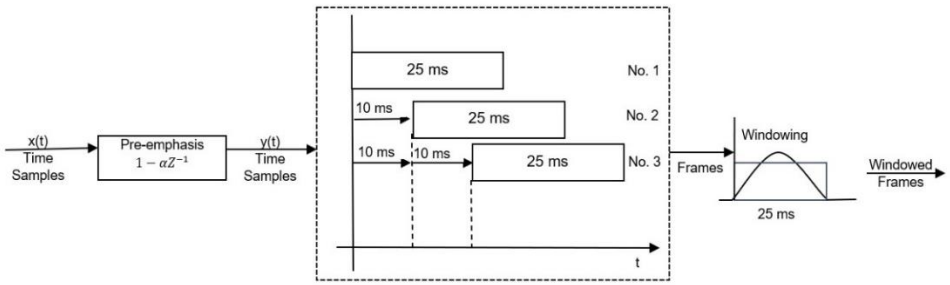

Fig. 3. Block diagram for Framing Analysis [6].

1) Pre-emphasis: The waveform is subjected to a high-pass filter. This emphasises higher frequencies and compensates for the attenuation of high frequencies during the human speech production process. A simple first-order high-pass filter with a typical co-efficient of 0.97 is used, i.e the filter function is,
$$y(t) = x(t) - 0.97\,x(t-1) \quad (1)$$
where x(t) is the input speech data and y(t) are the output.
2) Framing: The utterance's time-domain waveform is divided into overlapping fixed duration segments called frames. Typical frame durations range from 20 ms to 30 ms (usually 25 ms), every 10 ms, a frame is created (thus Consecutive 25 ms frames generated every 10 ms will overlap by 15 ms).
3) Windowing: A window function is applied to each frame. By tapering each frame at the beginning and end edges, the window function smoothes the effect of using a finite-sized segment for subsequent feature extraction. The Fourier Transform is used because most features are spectral in nature, and the window function's multiplicative effect in the time domain is convolutive in the spectral domain. A smoother and less distorted (by artefacts) spectrum is produced by a tapered window function. Without a specified window function, the framing operation produces a rectangular window effect, which produces undesirable spectral artefacts. The Hamming window function is the most common among the window functions used in FIR Digital filter design [6].

To encapsulate feature extraction, The standard steps for extracting speech features from a specific speech sequence are as follows: division of the sampled signal into 20-30 ms blocks, multiplication of the blocks by a window function (typically Hamming window), DFT transform calculation (typically using FFT), mel-scaling, and finally MFCC calculation [7].

B. MFCC Features

It is critical for speaker recognition to extract features from each frame that can capture speaker-specific characteristics. Many of these features have been studied in the literature [8]. LPCs (Linear Prediction Coefficients) have gotten a lot of attention [9] because they are directly derived from the speaker's speech production model. Perceptual Linear Prediction (PLP) coefficient [8], [10] are also used because they are based on human perceptual and auditory processing. However, spectral-based features, most commonly derived by direct application of the Fourier Transform, have gained popularity over the last two decades. According to research [10], the same features used in speech recognition are equally effective when used in speaker recognition. These characteristics are known as Mel-Frequency spaced Cepstral Coefficients (MFCCs), and their success stems from the use of perceptually based Mel-spaced filter bank processing of the Fourier Transform, as well as the robustness (to the environment) and flexibility that cepstral analysis can achieve. As shown in Fig. 4, MFCC features are derived.

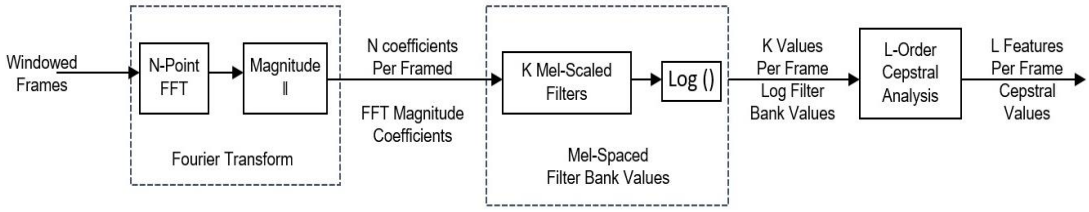

Fig. 4. Block diagram for MFCC feature vectors analysis [6].

III. Speaker Modeling Techniques

This section will go over various state of art speaker modeling techniques.

A. Vector Quantization (VQ)

First, the VQ model, also known as the centroid model, is one of the most basic text-independent speaker modeling techniques [3]. It was first used in speaker recognition in the 1980s. VQ, like GMM, is a generative classifier that estimates the feature distribution within each speaker. When combined with background model adaptation, VQ provides good accuracy [11].

The average quantization distortion is as follows:
Let the feature vectors of test utterances be denoted by $X = \{x_1, x_2, x_3, \ldots x_T\}$ and reference vector by $R = \{r_1, r_2, r_3, \ldots . r_k\}$

The average quantization distortion then becomes

$$D_Q(X,R) = \frac{1}{T}\sum_{t=1}^{T} min_{1\le k\le K}\, d(x_t, r_k) \quad (2)$$

where d (. , .) is the Euclidean distance defined as $\|x_t - r_k\|$

Note to remember that, $D_Q(X,R) \neq D_Q(R,X)$

The advantage of VQ is that it is a simple and efficient way to do speaker identification.

B. Gaussian Mixture Model (GMM)

The Gaussian Mixture Model (GMM) has developed itself as the standard classifier for text-independent speaker recognition over the last decade [12]. Because of its high recognition ability, the Gaussian Mixture Model (GMM) is frequently used in speaker verification[13]. The GMM's ability to form smooth approximations to arbitrarily shaped distributions, is one of its most powerful features. GMMs have distinct advantages over other modeling approaches, they can be trained quickly, scaled and updated to add new speakers with relative ease [14]. A Gaussian Mixture Model (GMM) is a parametric probability density function that is represented by the sum of Gaussian component densities. GMMs are commonly

used in biometric systems as a parametric model of the probability distribution of a continuous measurement of features.

The form of a GMM is a weighted sum of M component densities.

$$p(x|\lambda) = \sum_{i=1}^{M} w_i \, b_i(x) \quad (3)$$

where x is a dimensional random vector, $b_i(x)$, i=1,2……M, is the component densities and w_i, i=1,2,….,M, is the mixture weights.

The Gaussian Function can be defined of the form

$$b_i(x) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_i|^{1/2}} e^{\left\{-\frac{1}{2}(x-\mu_i)\cdot\Sigma_i^{-1}(x-\mu_i)\right\}} \quad (4)$$

with mean vector μ_i and covariance matrix $\sum_i$. The weight of the mixture satisfies the constraint that $\sum_{i=1}^{M} w_i = 1$[11].

The advantage of this approach is to provide a simple yet effective speaker representations which is computationally inexpensive and provides high recognition accuracy.

C. Universal Background Model (UBM)

When making an accept or reject decision in a speaker recognition system, a UBM or World Model represents general, person-independent, channel-independent feature characteristics that are compared to a model of speaker-specific feature characteristics. In this case, the UBM is a speaker independent GMM that has been trained on many speech samples to reflect general speech characteristics. The UBM is also used as a prior model in Maximum A Posteriori (MAP) parameter estimation when training the speaker-specific model [11].

The UBM is a wide GMM (1024 mixtures) that has been trained to describe the speaker-independent feature distribution. To train the UBM, simply pool all the speech data from an equal number of male and female speakers and run it through the Expectation-Maximization (EM) algorithm[15]. An effective method of performing speaker recognition is MAP adaptation, which integrates coupled target and background speaker model components. The ability to distinguish between regions of space that the GMM has learned from training speech is a major benefit of a fully coupled system. If there is no adaptation observation in the region near a mixture component, the mixture component will remain unadopted. However, because of applying adaptation, mixture components near the training observation will be adjusted to the speech data. Therefore, adapted regions will be more discriminative [16].

The main advantage of this approach is that a single speaker-independent model can be trained once for a particular task and then used for all hypothesized speakers in that task. It is also possible to use multiple background models tailored to specific sets of speakers.

D. Support Vector Machine (SVM)

A support vector machine (SVM) is a flexible discriminative classifier that has recently gained a lot of traction in the field of speaker recognition [3]. It is a two-class discrimination technique that entails locating a hyperplane to effectively separate the two classes under consideration. An SVM is a discriminative model that has been applied with spectral [17], [18], prosodic [19], [20], and high-level feature vectors [21] to determine the boundary between a speaker and a set of imposters. SVM can also be combined successfully with GMM to improve performance. SVM speaker recognition methods that are commonly used are based on comparing speech utterances using sequence kernels. In this case, the target speaker's utterances as well as a set of background speaker's utterances with impostor population characteristics are used to train a target model. A point in the SVM space is created for each speech sample from a target or background speaker.

SVM is a two-class classifier built from kernel function sums K(. , .)

$$F(x) = \sum_{i=1}^{N} \lambda_i \, t_i \, K(x, x_i) + d \quad (5)$$

where the t_i are ideal outputs, $\sum_{i=1}^{N} \lambda_i \, t_i = 0$ and $\lambda_i > 0$.

The vectors x_i are support vectors obtained through an optimization process from the training set. As shown in Fig. 5, the ideal outputs are either 1 or -1, depending on whether the corresponding support vector is in class 0 or class 1. A class decision is made based on whether the value, F(x), is greater than or less than a threshold [11].

The advantage of SVM is that it can be used for the data that is not regularly distributed and have unknown distribution.

E. Recognition Rates for Different Speaker Modeling Techniques

According to the experimental results in [11], as shown in Table I, VQ has the poorest performance when compared to other modeling techniques, with an EER value of 11.08 %. Fusions of GMM and SVM, on the other hand, improve performance by 0.7 % over a single SVM technique. Similarly, GMM-UBM outperforms traditional GMM, and SVM improves by about 2.8 % when compared to GMM-UBM. Above all, SVM outperforms its correspondence modeling techniques, GMM, by 3.72 percent.

F. Deep Neural Network (DNN)

An artificial neural network (ANN) with several hidden layers between the input and output layers is known as a deep neural network (DNN). These additional layers extract features from lower layers. This enables efficient modeling of complex data. DNNs are divided into two types: feed forward networks and recurrent neural networks. In recent years, DNNs have been used in the field of speech analysis; specifically in the domain of speaker identification, where they have proven to be far more effective than traditional techniques[22] [23]. A recent research has found that DNN outperforms MFCC in terms of efficiency [22]. Another study looked at DNN in a noisy and reverberant environment and found that it performed well [24].

A research published in [25] looked into the issue of speaker recognition in multi-talker speech with up to five simultaneous speakers. In addition to the commonly used

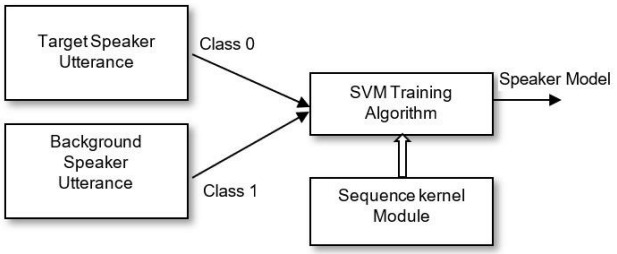

Fig. 5. Training module of SVM for speaker verification [11].

TABLE I: EER VALUES FOR THE SV SYSTEM MFCC AND PROSODIC FEATURES WITH DIFFERENT SPEAKER MODELING TECHNIQUES [11].

Speaker Modeling Techniques	EER %	Recognition Rate %
VQ	11.08	88.92
GMM	9.93	90.07
GMM-UBM	8.91	91.09
SVM	6.21	93.79
GMM-SVM	6.06	93.94

GMM-based approach, a deep learning-based approach is suggested to develop overlapping speaker identification (OSID) systems. Based on one dimensional convolutional neural network (1DCNN), multilayer perceptron (MLP), and GMM classifiers, the systems were designed in two types: two-stage OSID (T-OSID) and single-stage OSID (S-OSID). The experimental result in this research shows that the 1DCNN-based T-OSID system outperformed all other systems in each OSID scenario. On the evaluation dataset mixed at equal overlapping energy ratio (OER = 0dB), the 1DCNN-based T-OSID system achieved an accuracy of 98.55 percent with up to five simultaneous speakers. Furthermore, under more difficult experimental conditions that included high levels of noise (SNRs of 5dB and 0dB) and high OERs (5dB and 10dB) at the same time, the 1DCNN-based T-OSID system still achieved accuracies of greater than 90% [25].

The next section presents the most used application areas of speaker recognition as well as the standard modeling techniques used in these applications.

G. Comparison Between VQ And GMM

Fig. 6, depicts the most important results presented in [7], namely the overall recognition accuracy as a function of the number of centroids for VQ and the number of Gaussians for GMM. It can be shown that in both cases, text-dependent recognition accuracy is greater than 81.4%. The vector quantization algorithm was tested with 5 different numbers of centroids: 16, 24, 32, 48, and 64. As can be seen in some cases, the number of centroids has no effect on recognition accuracy, but it is increasing overall. Too many centroids increase computing time and can lead to algorithm overlearning, so the number of centroids has been limited to 32.

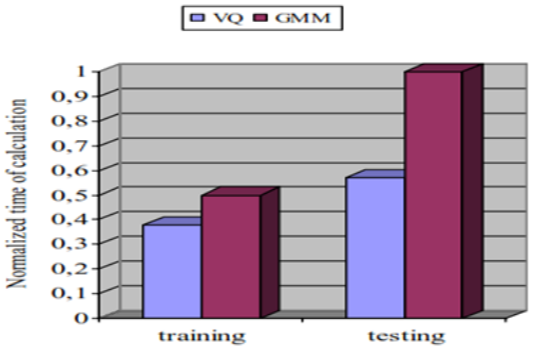

Fig. 6. Time of calculation [7].

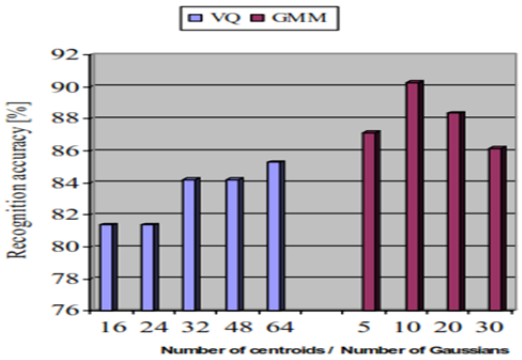

Fig. 7. Speaker recognition accuracy for VQ and GMM [7].

Fig. 7, depicts a comparison of the normalised time of the training and testing phases. In the case of GMM, the test phase is twice as large as the training phase and twice as large as VQ computing. When different numbers of testing files were used for each speaker (29 for VQ and 25 for GMM), the real testing time ratio GMM to VQ was 2. Parameterization and Euclidian distance computation take approximately 14 ms for the VQ algorithm and 28 ms for the GMM algorithm (Matlab environment v.7.8.9, computer efficiency by Matlab Bench Relative Speed=20) [7].

The resultant identification efficiency of more than 90% for GMM (10 Gaussians) and 84 % for VQ (32 centroids) demonstrates that a relatively simple vector quantization method works well for very short expression times of less than 3 seconds [7].

H. Comparison Between GMM, UBM And SVM

The results shown in Fig. 8, as illustrated in [6], demonstrate the GMM's sensitivity to the amount of training data and the number of mixture components. The best result obtained with the GMM was only 79.7% with 16 mixtures, and as more mixture components were added, rapid degradation was observed. Since the modeling techniques discussed so far vary in their working principles, they can be combined to increase the recognition rate even further.

With 128 mixtures, the GMM-UBM and GMM-SVM systems achieved a much better performance of around 95 percent. To put these findings into perspective, commercial biometric recognition applications require error rates of no more than 2% (ideally 1% or less)

depending on the number of speakers enrolled in the system [26]. Despite representing different classifier paradigms, the GMM-UBM and GMM-SVM achieved comparable performance based on the results presented this far. One possible explanation is that the SVM's supervector is based on the same UBM used by the GMM-UBM. Another possibility is that the data set used did not fully exploit any of the discriminant SVM classifier's advantages over a generative GMM-based system.

To investigate the variations in greater depth; the original feature sets and a reduced feature set (i.e., only a 13-dimension, rather than 39-dimension, MFCC Feature vector) were tested on the small training data set for 64, 128 and 256 Mixtures, with the results shown in Table II. When both the amount of training data and the number of features are reduced, the GMM-SVM achieves superior recognition of about 3% over the GMM-UBM in all cases, compared to when only the training data is limited. The results demonstrate an SVM classifier's potential superiority when presented with limited amounts of training data and/or reduced feature sets [6].

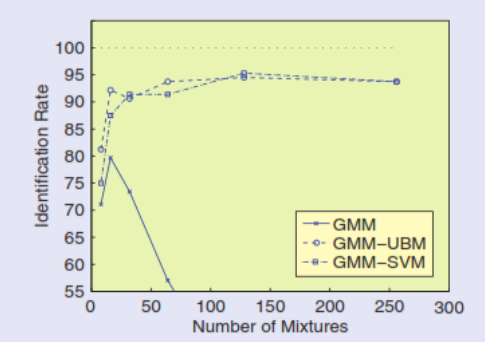

Fig. 8. Identification rates with three training utterance [6].

TABLE II: IDENTIFICATION RATES USING ORIGINAL FEATURES AND REDUCED FEATURES (WITHOUT THE TEMPORAL DERIVATIVES) BASED ON 3 UTTERANCES PER SPEAKER FOR TRAINING [6].

Mixtures	GMM-UBM		GMM-SVM	
	original	reduced	original	reduced
64	93.8	92.1	91.4	95.3
128	94.5	93.0	95.3	96.0
256	93.8	89.8	93.4	92.1

IV. SPEAKER RECOGNITION APPLICATIONS

A. Speaker Recognition for Authentication

The science of using a person's voice as a uniquely identifying biological characteristic to authenticate him is known as voice biometrics. Voice biometrics, also known as voice verification or speaker recognition, enables fast, frictionless, and highly secure access for a wide range of use cases, including call centres, mobile and online applications, chatbots, IoT devices, and physical access. During authentication, the system uses the same authentication utterance to identify who a speaker claims to be and to confirm whether the speaker is the claimed person, this is done using speaker recognition. It is more difficult to imitate and, in general, more convenient for users because they do not need to remember passwords or carry a physical token that can be easily lost or stolen. The authenticator is a part of the person.

Passwords are the most used method for protecting users' information. This method necessitates the user remembering his password for a long period of time. Furthermore, most users have multiple accounts, which means they must remember multiple passwords. Passwords can be forgotten or stolen as some people or accounts maybe hacked in some cases. Biometric Identification Systems are one solution to this annoyance. These systems rely on biometric characteristics of an individual that are unique to users, thus distinguishing them from one another.

There are many ways to accomplish this; one method is to use a speech recognition system as a gateway for security access control in order to be authorised for restricted services such as phone banking, voice mail, or access to database services, as detailed in [27].

This system's identification module employs Hidden Markov Model Toolkit (HTK). Hidden Markov models (HMM) for each enrolled speaker is created using this toolkit.

Another technique proposed in [28] is a hybrid speaker recognition system based on Mel Frequency Cepstrum Coefficient (MFCC) feature extraction and a combination of vector quantization (VQ) and Gaussian Mixture Modeling (GMM) for speaker modeling. This method recognises the speaker for both text dependent and text independent speech and employs relative indexes as confidence measures in the event of a contradiction in the recognition process by GMM and VQ. Based on the performance evaluation, the combination (GMM+VQ) outperforms individual models for speaker identification.

Voice recognition has several significant advantages over other methods of identity authentication:

- Because all phones have microphones, it is widely available for authentication on mobile phones.
- Low cost of integration into other devices such as automobiles and home appliances.
- Convenient and familiar to most users.
- Because it is contactless, it is less invasive and more hygienic.
- Beneficial for phone-based applications such as customer service.

The following are some disadvantages of voice recognition:

- It is not as precise as other biometric modalities (e.g., facial recognition).
- Live detection is required to ensure that a sample is from a live speaker rather than a recording.
- Background noise can influence sample quality and, as a result, matching performance.
- Not appropriate for all environments (e.g., noisy or public spaces).

B. Forensic Speaker Recognition

There has long been a desire to be able to recognise an individual solely by their speech. For years, judges, lawyers, detectives, and law enforcement officials have

desired to use forensic voice authentication to prosecute a suspect or validate a guilty or innocent judgment. Identifying a voice from forensic-quality samples is a difficult task for automated, semiautomatic, and human-based methods. The speech samples being compared may have been captured in various situations; for example, one sample may have been a screaming over the phone, while the other could have been a whisper in an interview room. In one or more of the samples, a speaker may be masking his or her voice, sick, or under the influence of narcotics, alcohol, or stress. The speech samples would almost certainly contain noise, be short, and lack sufficient appropriate speech content for comparison.

There are several approaches presented for this application, including the ones listed below.

- The GMM-UBM method, which is the most widely used in text-independent speaker recognition.
- Using a combination of GMM and SVM.

In the field of speech, discriminant classifiers based on support vector machines (SVM) were of great interest. A significant evolution in speaker recognition has been suggested, primarily by [18]. It employs a hybrid approach, combining the robustness of the GMM-UBM paradigm's statistical modeling with the discriminant power of SVMs. The GMM-UBM is used to model the training or testing data in this approach, known as GMM supervector SVM with linear kernel (GSL). In Research [29], a more detailed explanation of these two approaches is provided.

Automatic Speaker Recognition can also be used for Mobile Forensic Applications, where the GMM-UBM model is implemented, which is a state-of-the-art speaker recognition system that combines a GMM with a universal background model (UBM), as detailed in [30].

C. Speaker Recognition for Surveillance (Law Enforcement)

Speaker recognition technology has two primary applications. The verification of cooperative speakers on a trial basis is the most frequently considered. The other is for monitoring a large set of speech samples in order to locate a specific speaker of interest. The obvious applications are in law enforcement and intelligence, but areas such as news indexing are also relevant [31].

Surveillance systems do not need to make difficult decisions, and in some cases, it may be unwise to do so. A better application might be to use it as a tool for prioritising samples for further examination by experts. The surveillance system is assumed to be operating on a large set of samples, with the task of interest being to locate the speaker of interest. While this may seem to be arbitrary limitation, it accurately models a form of surveillance application in which the primary objective is to find a single target, regardless of the number of actual targets. This could be because, once a single target has been identified, the associated collateral information allows easier detection of all remaining targets [32].

Today's cutting-edge speaker recognition technologies employ Gaussian mixture models (GMMs), which exhibit this type of behaviour. It is assumed that true speaker scores, like false speaker scores, are distributed as univariate Gaussian variables.

Most surveillance systems necessitate some level of human intervention in the results. As a result, forcing the system's automated components to make hard accept/reject decisions may not be necessary. Instead of focusing on sample prioritisation, automated systems can achieve significant improvements in performance, as measured by false alarms per target detection for truncated queues. Research [32] provides a model for the performance of systems with Gaussian output score distributions for target speakers and false speakers. This model was shown to be a good match for the performance of a real-world speaker recognition system.

Another technique which described the speaker recognition problem in relation to the complex surveillance system is presented in [33], where a proposed system extension enables identified the precise identity or at least the gender of the suspect by the captured voice analysis. The solution is based on a text-independent approach that uses Mel Frequency Cepstral Coefficients and fundamental frequency to extract the identity from a voice signal. This system expansion could aid in the elimination of vandalism and the elucidation of crimes [33].

Text-independent speaker recognition is a new feature of the acoustic event detection (AED) system EAR-TUKE [34]. It detects shooting and breaking glass sounds and is built to work in an outdoor setting. The MFC front-end (Mel-Frequency Cepstrum), the Hidden Markov Model classification approach, and the modified Viterbi-based decoding algorithm form the foundation of the AED functional prototype [33].

D. Speech Recognition (Speech Data Management)

As shown in Fig. 9, the state-of-the-art approach models human speech development and recognition across four stages: text generation, speech production, acoustic processing, and linguistic decoding [35].

The Speech Recognition approach aims to recognise text from speech utterances, which can be more useful for people who are deaf or hard of hearing. Speech recognition techniques such as Support Vector Machine (SVM) and Hidden Markov Model (HMM) are commonly used.

In research [36], modeling techniques such as SVM and HMM were used to model each individual word, resulting in 620 models that were trained on the system. To determine the semantic representation of the test input speech, each isolated word segment from the test sentence is compared to these models. The system's performance is evaluated for computer domain words, and the system achieves an accuracy of 91.46 percent for SVM and 98.92 percent for HMM. Based on the extensive analysis, HMM outperforms other modeling techniques such as SVM.

Over the last two decades, automatic speech recognition (ASR) technology has advanced significantly. The most well-known systems today can recognise phone quality and spontaneous speech, while earlier systems could only recognise isolated words. Basic acoustic Modeling and feature extraction techniques, on the other hand, have not evolved, and the ASR system is still far from human-like performance in real-world scenarios. Most advancements in ASR systems have been made in the areas of pre-processing, feature extraction, language modeling, and model adaptation of such HMM/ GMM systems. As a result, it has been a focus of research.

In research [37], an HMM/GMM hybrid ASR system is presented, with HMM serving as a model of the sequential structure of speech signals and each HMM state employing a GMM to model the acoustic characteristics of sound units. The most common spectral representation is a set of mel frequency cepstral coefficients (MFCCs), or perceptual linear prediction (PLP) derived from a window of about 20 ms of speech overlapped by about 10 ms frame, with each frame of coefficients being augmented with differences and differences of differences with nearby frames. The GMM can be thought of as a cross between parametric and nonparametric density models. It has a structure and parameters, just like a parametric model, that control density behaviour in predictable ways. It has several degrees of freedom, much like a non-parametric model, allowing for arbitrary density modeling. To model the distribution of feature vectors for a given state, each state of HMMs is represented by a GMM.

E. Speaker Tracking

The process of tracking speakers in continuous audio streams entails several processing tasks. Thus, it is classified as a multistage process. The components for audio segmentation, speech detection, speaker clustering, and speaker identification are the main building blocks of a system of speaker tracking. The first three processes seek to identify homogeneous regions in continuous audio streams that belong to a single speaker and to connect each region of the same speaker.

The task of organising audio data in this manner is known as speaker diarization, and it is important in many speech-processing applications [38].

In all cases, speech detection was accomplished by classifying each segment of an audio stream as speech or non-speech using the GMM that produced the highest likelihood from the given data, whereas in the phoneme-recognition and fusion cases, only two GMMs were used. The speech segments that were detected were then transferred to a speaker clustering module, while the non-speech segments were discarded from further processing [38].

A speaker-identification component was adapted from a speaker-verification system originally designed to detect speakers in conversational telephone speech, [39]. The basic Gaussian Mixture Model – Universal Background Model (GMM-UBM) method was used to create the speaker verification system [38].

Two approaches for detecting and tracking speakers in multi-speaker audio are described in [40]. The core speaker recognition engine in both approaches is an adapted Gaussian mixture model, universal background model (GMM-UBM) speaker detection system.

Individual log-likelihood ratio scores generated on a frame-by-frame basis by the GMM-UBM system are used in the first approach to partition the speech file into speaker homogeneous regions and then create scores for these regions.

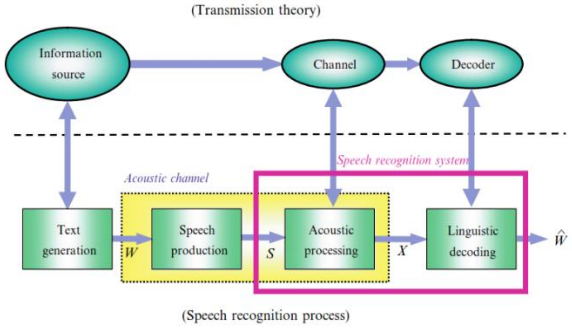

Fig. 9. Structure of the state-of-the-art speech recognition system [37].

This approach is known as internal segmentation. The other method partitions the speech file into speaker homogeneous regions using an external segmentation algorithm based on blind clustering. As in the case of single-speaker recognition, the adapted GMM-UBM system scores each of these regions. For both detection and tracking, it is demonstrated that the external segmentation system outperforms the internal segmentation system.

F. Security

VOIP (Voice over Internet Protocol) is a promising technology that has been predicted to be the future of voice communication. As the use of internet-enabled devices becomes more widespread, the demand for speaker recognition systems over VoIP networks will grow. Research [41] analyses and investigates the relationship between the amount of VoIP speech data, optimal speaker model size, and performance in text-independent speaker identification over VoIP networks. This research is carried out using a cutting-edge speaker identification system that employs Mel frequency cepstral coefficients (MFCC) as acoustic features and Gaussian mixture models (GMM) for speaker modeling. The most fascinating conclusion that can be drawn from this research is the critical importance of speaker model size for system robustness.

The majority of voice over IP (VoIP) traffic is encrypted before being transmitted over the Internet. Since traditional speaker recognition methods are restricted to unencrypted speech communications, tracing the identity of perpetrators during forensic investigations is a difficult task.

Among the significant work done in the area of speaker recognition, the Gaussian mixture model universal background model (GMM-UBM) [42] and the mixed GMM-UBM and SVM technique are commonly used in text-independent speaker recognition problems,

particularly in speaker verification or source confirmation disputes. The mixed GMM-UBM and SVM approach combines the modeling efficacy of Gaussian mixtures with the discriminative power of SVMs, resulting in a significant improvement in identification accuracy. In the case of speaker identification, accuracy is measured simply as the ratio of correctly identified speech segments to the total number of segments in a group of speakers. This accuracy metric is heavily influenced by the potential number of suspects; as the population size grows, the accuracy decreases. As a two-class classification problem, speaker verification can result in two types of errors: false rejection (rejecting a valid speaker) and false acceptance (accepting a valid speaker) (accepting an invalid speaker).

In the research paper [43] various proposed techniques for speaker identification and verification from encrypted VoIP conversations is introduced, where experimental results demonstrate that these techniques can correctly identify the actual speaker 70–75 percent of the time among a group of ten possible suspects. Table III summarizes the application and the relevant modeling technique that was used with it, as shown in the preceding sections. Table III summarizes the application and the relevant modeling technique that was used with it, as shown in the preceding sections.

TABLE III: APPLICATIONS OF SPEAKER RECOGNITION AND ITS RELEVANT TECHNOLOGY

Application	Technology (modeling technique)
Speaker recognition for Authentication	• Hidden Markov models (HMM) • Combination (GMM+VQ)
Forensic Speaker Recognition	• The GMM-UBM approach • The mixed GMM and SVM approach
Speaker recognition for surveillance (law enforcement)	• Gaussian mixture models (GMMs) • Acoustic Event Detection (AED) system
Speech recognition (Speech data management)	• Support Vector Machine (SVM) • Hidden Markov Model (HMM)
Speaker Tracking	• (GMM-UBM) approach (internal segmentation) • (GMM-UBM) approach (external segmentation)
Security	• The Gaussian mixture model universal background model (GMM-UBM) • Mixed GMM-UBM and SVM approach

V. CONCLUSION AND FUTURE TRENDS

For decades, researchers' focus has been raised to develop reliable speaker recognition (SR) for security and authentication applications; however, nowadays, we live in a new era. The new normal way for our communication is the VOIP calls, especially for business and governmental work, which created a high demand for an application for speaker recognition and noise cancellation. The identification of individuals can help with enhancing the quality of the calls by isolating the voice of the concerned person from any other human or non-human voice around him. This paper gives an overview of the technologies for speaker recognition such as Gaussian Mixture Model (GMM), Support Vector Machine (SVM),

Universal Background Model (UBM), and Deep Neural Network (DNN), which are used to recognize the speaker voice for authentication purposes. Focus is also directed to the various applications of speaker recognition. However, many studies may concentrate on speaker recognition technologies or modeling techniques, and other studies focus on the applications. This paper tried to link this gap and presented different modeling techniques that were used in each application of speaker recognition to improve performance.

To conclude, the fundamentals of speaker recognition was briefly introduced, including feature extraction and modeling techniques. It is noteworthy to observe that while MFCC is considered as one of the most widely used feature extraction techniques for SR, the GMM model is widely implemented for print matching of speaker voice. Further, this paper presented various technologies that have been used for different SR applications in several fields.

The authors observe from the relation between SR techniques and applications that GMM model is mainly the traditional model that is used in most of the applications even alone or combined with other modeling techniques.

We believe that there is an enormous potential for speaker recognition technology in multimedia and biometric applications. The existing technologies and applications aimed to identify the speaker voice for authentication regardless of the speech and/or distinguish the human voice from the non-human. These techniques have proven their efficiency for authentication applications and recorded voices. However, real-time calls

are still a research gap that needs to be studied intensively. The aim is to deliver a secure and clear targeted speaker voice over the call and cancel the surrounding environment on real-time VOIP calls. These challenges motivate further research and investment in some of the following important directions:

- The speaker recognition using Deep Neural Network from other human voices and non-human audio.
- A real-time artificial intelligent algorithm for speaker identification.
- Appling speaker voice separation algorithm for VOIP calls.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

In this paper, all authors have their contributions. Amira Mohamed prepared and analyzed the data and wrote the paper; Amira Eltokhy proofread the research; Abdelhalim Zekry reviewed the paper; all authors had approved the final version.

References

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. CRC Press, 2013.

[2] S. Bunrit, T. Inkian, N. Kerdprasop, and K. Kerdprasop, "Text-independent speaker identification using deep learning model of convolution neural network," *Int. J. Mach. Learn. Comput.*, vol. 9, no. 2, pp. 143–148, 2019.

[3] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Commun.*, vol. 52, no. 1, pp. 12–40, 2010.

[4] S. Furui, "Speaker recognition," *Scholarpedia*, vol. 3, no. 4, p. 3715, 2008.

[5] N. Singh and A. Agrawal, "Principle and applications of speaker recognition security system," no. June, 2018.

[6] R. Togneri and D. Pullella, "An overview of speaker identification: Accuracy and robustness issues," *IEEE Circuits Syst. Mag.*, vol. 11, no. 2, pp. 23–61, 2011.

[7] M. H. Farouk, "Speaker Recognition," *SpringerBriefs Speech Technol.*, pp. 33–35, 2014.

[8] R. J. Mammone, X. Zhang, and R. P. Ramachandran, "Robust speaker recognition: A feature-based approach," *IEEE Signal Process. Mag.*, vol. 13, no. 5, pp. 58–71, 1996.

[9] B. Atal, "Wave for automatic speaker identification and verification," *Jasa*, vol. 55, pp. 1304–1312, 1974.

[10] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," vol. 2, no. 4, pp. 639–643, 1994.

[11] K. Sarmah, "Comparison studies of speaker modeling techniques in speaker verification system," *Int. J. Sci. Res. Comput. Sci. Eng.*, vol. 5, no. 5, pp. 75–82, 2017.

[12] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *ESCA Work. Autom. Speak. Recognition, Identification, Verif. ASRIV 1994*, vol. 17, pp. 27–30, 2019.

[13] N. Malayath, H. Hermansky, S. Kajarekar, and B. Yegnanarayana, "Data-driven temporal filters and alternatives to GMM in speaker verification," *Digit. Signal Process. A Rev. J.*, vol. 10, no. 1, pp. 55–74, 2000.

[14] A. Fazel and S. Chakrabartty, "An overview of statistical pattern recognition techniques for speaker verification," *IEEE Circuits Syst. Mag.*, vol. 11, no. 2, pp. 62–81, 2011.

[15] D. A. Reynolds, "Gaussian mixture models," *Encycl. Biometrics.*, 2008.

[16] J. Pelecanos, R. Vogt, and S. Sridharan, "A study on standard and iterative MAP adaptation for speaker recognition," in *Proc. Int. Conf. Speech Sci. Technol.*, pp. 190–195, 2002,

[17] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Comput. Speech Lang.*, vol. 20, no. 2-3 SPEC. ISS., pp. 210–229, 2006.

[18] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal* Process. *Lett.*, vol. 13, no. 5, pp. 308–311, 2006.

[19] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling prosodic feature sequences for speaker recognition," *Speech Commun.*, vol. 46, no. 3–4, pp. 455–472, 2005.

[20] L. Ferrer, E. Shriberg, S. Kajarekar, and K. Sönmez, "Parameterization of prosodic feature distributions for SVM modeling in speaker recognition," in *Proc. ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 4, pp. 2–5, 2007.

[21] W. M. Campbell, J. P. Campbell, D. A. Reynolds, D. A. Jones, and T. R. Leek, "Phonetic speaker recognition with support vector machines," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004.

[22] B. L´EON and P. H. Y. Lecun, "Gradient-Based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[23] S. S. Tirumala, "Deep learning using unconventional paradigms," in *Deep Learning: Fundamentals, Methods and Applications*, 2016.

[24] S. S. Tirumala and S. R. Shahamiri, "A review on deep learning approaches in speaker identification," in *Proc. ACM Int. Conf. Proceeding Ser.*, no. November, 2016, pp. 142–147.

[25] V. T. Tran and W. H. Tsai, "Speaker identification in multi-talker overlapping speech using neural networks," *IEEE Access*, vol. 8, pp. 134868–134879, 2020.

[26] A. K. Jain, A. Ross, and S. Prabhakar, "An Introduction to Biometric Recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 4–20, 2004.

[27] L. Dovydaitis, T. Rasymas, and V. Rudžionis, "Speaker authentication system based on voice biometrics and speech recognition," *Lect. Notes Bus. Inf. Process.*, vol. 263, pp. 79–84, 2017.

[28] D. Desai and M. Joshi, "Speaker recognition using MFCC and hybrid model of VQ and GMM," *Adv. Intell. Syst. Comput.*, vol. 235, pp. V–VI, 2014.

[29] J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J. F. Bonastre, and D. Matrouf, "Forensic speaker recognition," *IEEE Signal Process. Mag.*, vol. 26, no. 2, pp. 95–103, 2009.
[30] M. Algabri, H. Mathkour, M. A. Bencherif, M. Alsulaiman, and M. A. Mekhtiche, "Automatic speaker recognition for mobile forensic applications," *Mob. Inf. Syst.*, vol. 2017, 2017.
[31] A. Albiol, L. Torres, and E. J. Delp, "The indexing of persons in news sequences using audio-visual data," in *Proc. ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 3, pp. 137–140, 2003.
[32] P. J. Barger and S. Sridharan, "On the performance and use of speaker recognition systems for surveillance," in *Proc. - IEEE Int. Conf. Video Signal Based Surveill. 2006, AVSS 2006*, 2006.
[33] E. Kiktova and J. Juhar, "Speaker recognition for surveillance application," *J. Electr. Electron. Eng.*, vol. 8, no. 2, pp. 19–22, 2015.
[34] M. Lojka, M. Pleva, E. Kiktová, J. Juhár, and A. Čižmár, "EAR-TUKE: The acoustic event detection system," *Commun. Comput. Inf. Sci.*, vol. 429, no. June, pp. 137–148, 2014.
[35] S. Furui, "Speech and speaker recognition evaluation," *Eval. Text Speech Syst.*, pp. 1–27, 2007.
[36] S. Ananthi and P. Dhanalakshmi, "SVM and HMM modeling techniques for speech recognition using LPCC and MFCC features," *Adv. Intell. Syst. Comput.*, vol. 327, pp. 519–526, 2014.
[37] M. Sarma and K. K. Sarma, "Acoustic modeling of speech signal using artificial neural network: A review of techniques and current trends," *Intell. Appl. Heterog. Syst. Model. Des.*, no. June, pp. 282–299, 2015.
[38] J. Zibert, B. Vesnicer, and F. Miheliç, "Development of a speaker diarization system for speaker tracking in audio broadcast news: A case study," *J. Comput. Inf. Technol.*, vol. 16, no. 3, pp. 183–195, 2008.
[39] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, "NIST speaker recognition evaluation - overview, methodology, systems, results, perspective," *Speech Commun.*, vol. 31, no. 2, pp. 225–254, 2000.
[40] R. B. Dunn, D. A. Reynolds, and T. F. Quatieri, "Approaches to speaker detection and tracking in conversational speech," *Digit. Signal Process. A Rev. J.*, vol. 10, no. 1, pp. 93–112, 2000.
[41] B. Ayoub, K. Jamal, and Z. Arsalane, "Investigation of the relation between amount of VoIP speech data and perrformance in speaker identification taskover VoIP Networks," in *Proc. World Congr. Inf. Technol. Comput. Appl. WCITCA 2015*, 2015.
[42] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, 1995.
[43] L. A. Khan, M. S. Baig, and A. M. Youssef, "Speaker recognition from encrypted VoIP communications," *Digit. Investig.*, vol. 7, no. 1–2, pp. 65–73, 2010.

Amira A. Mohamed received the B.Sc. in Electronics and Electrical Communications and the M.Sc. degree (2017) in digital signal processing from Ain Shams University (ASU), Cairo, Egypt. She worked as an academic staff member at MSA and BUC, Egypt. Currently, she is an assistant lecturer at BUC and a Ph.D. student at ASU since 2019. Her research interests include communication engineering, signal processing and Machine learning.

Amira Eltokhy received the M.S. degree in wireless mobile communication systems and the Ph.D. degree in microwave engineering from the University of Greenwich, London, U.K., in 2015 and 2019, respectively. She was a Post-Doctoral Associate with Bangor University, Bangor, U.K., from 2019 to 2020. Currently, she is a lecturer at MSA University, Egypt. Her research interests include communication systems and biomedical engineering, in addition to artificial intelligence integration in both fields.

Abdelhalim A. Zekry is a professor of electronics at faculty of Engineering, Ain Shams University, Egypt. He worked as a staff member on several universities. He published more than 300 papers. He also supervised more than 110 Master thesis and 40 Doctorate. Prof. Zekry focuses his research programs on the field of microelectronics and electronic applications including communications and photovoltaics. He got several prizes for his outstanding research and teaching performance.