Al-Rafidain Engineering Journal (AREJ) Vol.29, No.2, September 2024, pp. 32-44

Methods and Techniques for Speaker Recognition: A Review

Abdalem A. Rasheed
alem12@uomosul.edu.iq

Mohammad Tariq Yaseen
mtyaseen@uomosul.edu.iq

Marwan A. Abdulhameed
marwanhajali@uomosul.edu.iq

Department of Electrical Engineering, College of Engineering, University of Mosul, Mosul, Iraq

Received: November 12th, 2023 Received in revised form: December 24th, 2023 Accepted: january 6th, 2024

ABSTRACT

An identity verification and identification system based on a person's distinctive vocal characteristics is known as speaker recognition. This paper sheds light on the evolution of speaker recognition systems from the earliest days of computers to the most recent innovations. Voice represents the behavior biometric that communicates details about a person's features, ranging from the speaker's age, gender, and ethnicity. The field of speaker recognition focuses on identifying individuals by their voices. Even though speaker recognition has been the subject of research for the past eight decades. Applications such as the Internet of Things (IoT), smart homes, and smart gadgets have made their use fashionable in the modern era. The speaker recognition field is briefly discussed in this work with an outline of its modeling methodology and various feature extraction strategies across multiple languages. The aim of this speaker recognition literature is to advance academic knowledge of speaker recognition.

Keywords:

Speaker recognition; Biometric; MFCC; GMM; Feature extraction.

https://rengj.mosuljournals.com
Email: alrafidain_engjournal2@uomosul.edu.iq

==

1. INTRODUCTION

A person can be recognized or verified using speaker recognition (SR), a technique that analyzes a person's speech characteristics. Human vocal cords produce voice, which the auditory system hears. The source and destination of the voice are the foundation of speaker recognition technology. While other speaker recognition algorithms simulate the human auditory system to obtain features heard in the ears, some focus on the pitch and frequency of the voice. Text-dependent and text-independent categories were used to classify the speaker identification based on whether they require reading a specific text to obtain voice information. Speaker recognition encompasses speaker identification and speaker verification[1].

2. FIRST SPEAKER RECOGNITION SYSTEM

Before a kidnapping and murder case in 1932, no academic study on SR had been done. To satisfy the suspect, Bruno Hauptmann, more than two years later, Charles Lindbergh, the victim's parent, by chance overheard the voice of the criminal close to where he was instructed to lay the ransom[2]. The first study on the validity of eyewitness testimony was started by Frances McGehee as a result of this legal case [3]. The study of the SR system has been continued, with data from McGehee's earlier work[4], [5] serving as a foundation.

The earliest reports of SR research date back to the 1960s. Using the spectrogram approach, Kersta [6] conducted the first study on speaker identification in 1962, and Li, et al. published the first study on speaker verification in 1966[7]. A spectrogram is a graphic representation of the size of the spectral properties that change over time, showing the relation between the frequency and the spoken signal energy related to time. Before that, in 1947, Bell laboratories' physicist wrote an article about voice identification [5]. The person's voice production system was modeled physically by Gunnar Fant in 1960 [8]. This approach offers the conceptual basis for explaining the speech processing for speaker and speech recognition. Fig. 1 depicts the timeline of significant developments in speaker recognition that have had a high influence on the development of the speaker recognition domain.

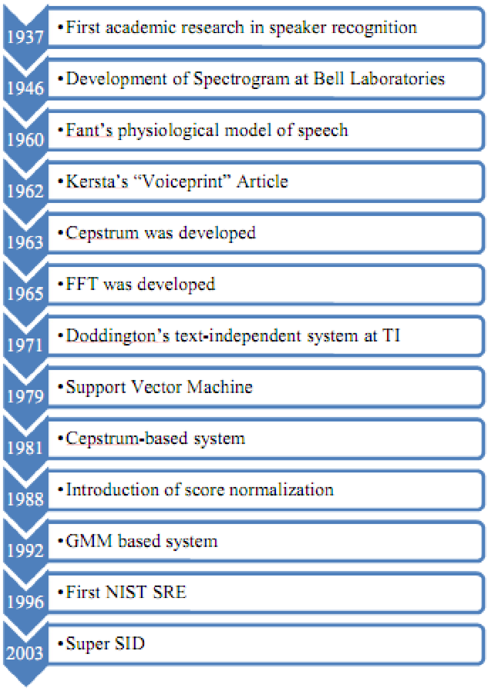
1937
•First academic research in speaker recognition
1946
•Development of Spectrogram at Bell Laboratories
1960
•Fant's physiological model of speech
1962
•Kersta's "Voiceprint" Article
1963
•Cepstrum was developed
1965
•FFT was developed
1971
•Doddington's text-independent system at TI
1979
•Support Vector Machine
1981
•Cepstrum-based system
1988
•Introduction of score normalization
1992
•GMM based system
1996
•First NIST SRE
2003
•Super SID

Fig. 1 Major development in speaker recognition across time [9], [10].

During the same period, innovations in the domain of computers were also made, which helped to solve numerous implementation issues with continuous and discrete words. The implementation of the FFT (Fast Fourier Transform)[9] was reported by Tukey and Cooly in 1965[10]; this paper presented a technique for frequency domain signal analysis in the computer. Tukey, Healy, and Bogert released a discus in 1963 titled "The Quefrency Analysis of the Time Series for Echos: Cepstrum, Pseudo-Auto-Covariance, Cross-Cepstrum, and Saphe Cracking" on the topic of seismic signal echoes[11]. By considering the spectrum log magnitude, which represents the relation between frequency and time, it provided a way of sound detection.

Michael Noll [12] first suggested using the cepstrum to ascertain the transmission pitch in 1969. Ronald Schafer joined the research, which resulted in the complex cepstrum, or the magnitude spectrum of the Fourier transform. To model speech using Noll's pitch detection technique, Schafer used cepstral analysis [13], [14]. Subsequently, speaker recognition systems extensively used the developed cepstral speech model.

In 1974 [23]–[26], Atal proposed Linear Prediction (LP) dependent characteristics for Automatic speaker recognition (ASR), including the impulse response function, Linear prediction cepstral coefficients LPCC, and autocorrelation function [15]–[18]. The cepstral data found and became the first important among all LP-based features, according to Atal's research [15]. The block diagram for the recognition regime is shown in Fig. 2.

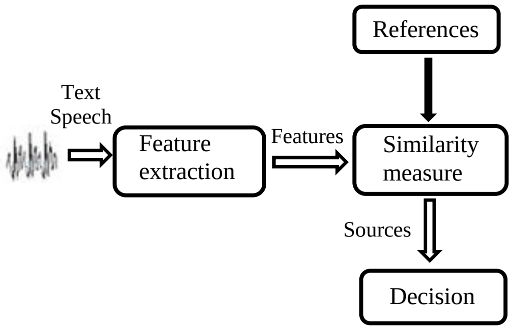

Fig. 2 Block diagram for the recognition system [20]

Since 1980, many speaker recognition systems for the reliant mode have been made available.

3. SPEAKER RECOGNITION TYPES:

Speaker recognition is a difficult task since different speech signals are produced during training and testing sessions by various circumstances, such as voice changes brought on by aging, illness, speaking rates, etc. [19], as shown in Fig.3.

There are several categories of speaker recognition, including: Identification, Verification, Segmontation, Clustering, Detection, and Diarization. [20], [21].

In addition, the speaker recognition that is classified into text-dependent and text-independent can also be divided into closed and open set systems [22], [23].

The exact text is said in a text-dependent system throughout both the stages of testing and training; however, in the text-independent system, there is no restriction on the text that is uttered, which the speakers find more practical[23]–[25]. The text-dependent system needs a short time for the procedure of training where it applies a certain set of input signals. In contrast, because the algorithm aims to identify the speaker by converting audio into distinctive without considering what is being spoken, the training stage of the text-independent system is more extended [26], [27]. Fig.3 shows the types of speaker recognition.

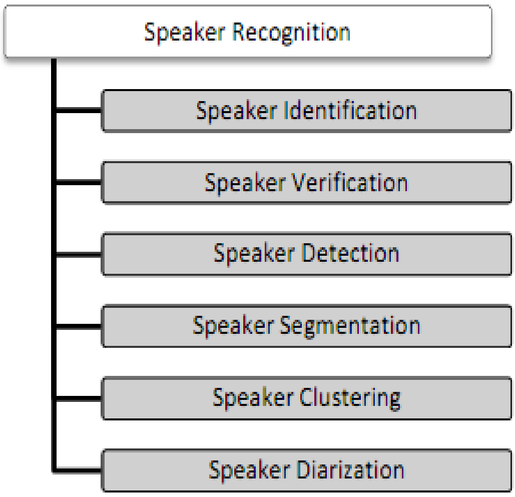

Fig. 3 Types of speaker recognition. [24] ,[29]

3.1 Speaker Identification:

In this type of recognition, the system uses the speech from unidentified speakers to identify which enrolled speaker most closely resembles the speech. This type of speaker identification algorithm selects the matching speaker from among the speakers who have signed up, and it is possible that the unknown person is different from those chosen by this kind [28]. Because of this, speaker identification is commonly joined with speaker verification in several systems [29].

3.2 Speaker Verification:

The purpose of the speaker verification is to make sure that an input utterance matches the identity that has been asserted. Determine whether the unfamiliar voice is that of a particular reference speaker. Then, accept the intended person or reject the liar [30]-[32].

3.3 Speaker Detection:

In speaker detection, the target person with the testing speeches is supplied to the system. Then the system will accurately label and discriminate the concerned person's talks [29], [34].

3.4 Speaker segmentation:

In this type of recognition, many inputs with the multi-speaker is present to the system. Then the system locates the points when the speaker varies. When the speaker's information is ready, produce models for every speaker. On the other hand, the system is known as blind speaker recognition [31], [33].

3.5 Speaker clustering:

The speaker clustering aim is to accurately group several talks offered to the system[31].

3.6 Speaker Diarization:

In the case of speaker diarization, the stream is offered to the system. Then, the system must locate the speaker who is said at each duration of the stream. This aim can be represented by constructing the stream segmentation aim joined (followed) with the clustering one. Then, if the information is ready for the system, the models will be assembled. Therefore, the task is known as model-based speaker recognition [31].

4. METHODS OF SPEAKER RECOGNITION

In the past fifteen years, there have been substantial advancements in each recognition system component, including characteristic choice characteristics, categorization characteristics, modeling characteristics, and making decisions. The improvements in the different speaker recognition areas contributed to its transformation from a purely academic activity to a practical fact.

4.1 Low and High Level

Short-term (10-20ms) voice features are represented by the low level, which has been the favored feature for most SR employment. However, the low-level method eliminates other distinguishable details in a speaker's speech. Pitch is one example of a low-level characteristic (e.g., the duration of silence between uttered words). The high-level features carried valuable information[15]. Early studies attempted to take advantage of this, but their results were restricted successes. With the introduction of the cepstram, low-level analysis once again becomes the focus of research[34].

4.2 Hidden Markov Model (HMM)

The HMM method represents a significant way for vocal modeling for speaker systems [35]–[37]. The ability of this model to analyze speech phenomena and its accuracy in real-world speech recognition systems are the primary factors in its success. The HMM's convergent and dependable parameter training process is another key feature. The representation of spoken utterances is the sequence of nonstationary vectors of features. As a result, segmenting the speech sequence into stationary states is necessary to achieve statistical calculation of the sequence of speech.

4.3 Vector Quantization (VQ)

A data classification technique called Vector Quantization (VQ) was created in 1979[38]. The notion of utilizing VQ in speaker recognition originates from its successful application in the recognition of hand-written digits[39]. The VQ technique process to organize the data coming from an accredited one or a fraud. By utilizing an optimized non-linear decision limiting, this method's capacity is of interest to reduce the rate of error of wrong rejection and wrong acceptance.

4.4 Dynamic Temporal Warping (DTW)

This approach is an automated speech recognition (ASR) technique based on template matching. This method matches the parameters of words against those of a single referenc template. Dynamic temporal warping (DTW) is used to adjust and deduce the similarity degree between the speaker role model and the sample sentence. Where this method takes a long time to process, and the system occupies a large memory[40].

4.5 Deep Neural Network (DNN)

The DNN model is a flexible network input layer. Where the input layers are diverse, the DNN model can add other demand characteristics that may assist in identifying the performance of the issued users. Then the DNN model can solve the limitations of factorization of matrix due to the versatile network input layer[41- [43].

4.6 Time Delay Neural Network (TDNN)

In voice recognition software, the acoustic model—which transforms the auditory signal into a phonetic representation—often employs the TDNN. The aim of the TDNN method is to categorize the patterns with shift-invariance and condition of the model at each network by applying the design of a multilayer artificial neural network. [44].

4.7 Gaussian Mixture Model (GMM)

Reynolds' doctoral research in 1992 focused on using Gaussian mixture models to model voice features for SR. His contributions helped SR adopt a new worldview[45], [46]. With a significant decrease in processing resources, the GMM performs as a preferable method. The GMM alone represents a substantial advance in recognizing technology. However, there have been a number of improvements made over the straightforward multivariate Gaussian mixture models. The Universal Background Model (UBM) was one of the most noteworthy improvements [46, 47]. In addition to modeling the voice of the person and estimating the probability that the person was the authenticated user, it was suggested to use a group of individuals who were not the authenticated user.

This made this model able to apply likelihood ratios and the Bayesian theory. [48]. For that specific system, the characteristics distribution of speaker-independent is represented by the GMM-UBM. Therefore, it is more likely that a worker is certified if their test utterance closely resembles the authenticated training data.

4.8 Discrete Wavelet Transform (DWT):

DWT frequently outperforms Fourier Transform-based parameterizations because it accurately represents the signal in both the frequency and time frame domains[49]. Karl [50] explored the parametric of the Czech language with a classifying approach. He shows that the wavelet configuration gave a high identification rate and reduced training time. As a result, current research efforts concentrate on applying the Wavelet Transform to many aspects of autonomous speech processing.

4.9 Continuous Wavelet Transform CWT

A signal's wavelet decomposition is accomplished via the Continuous Wavelet Transform (CWT). Small oscillations known as wavelets are highly concentrated in the time range. The basis functions of the CWT are scaled and shifted variations of the time-localized origen wavelet (mother wavelet), while the FFT distributes a signal into sines and cosines of indefinite length. In FFT, the time-localization information will be lost. The time-frequency description that was produced by the CWT provides superior time and frequency [51].

4.10 Neural Network Classifiers NNC

The NNC contains cells that are arranged in layers. These layers transform the vector of input to the output. Every unit receives an input processed with a particular operation, which is frequently nonlinear, and then sends the results to the following layer [52].

4.11 Support Vector Machine (SVM)

Support Vector Machine (SVM) represents a method that uses each phrase trained to the system and modeled. To identify the semantic description of the test input voice, each segment of the isolated word is compared with against these models [53].

5. FEATURE EXTRACTION TECHNIQUES:

5.1 Mel Frequency Cepstral Coefficients (MFCC)

Any machine learning method's performance is greatly affected by feature extraction and representation. The Mel Frequency Cepstrum Coefficient (MFCC), popular in many domains, was created to model the characteristics of audio signals. “The shape of the vocal tract manifests itself in the envelope of the short time power spectrum, and the job of MFCCs is to accurately represent this envelope.” [19], [54], [55].

5.2 Linear Predictive Coding (LPC)

The approach of LPC, which is frequently employed in audio signal processing, is used to illustrate the envelope of the signal of a compressed digital voice signal. Where the most popular technique for voice coding and speech synthesis is LPC[56]

5.3 Linear Predictive Cepstral Coefficient (LPCC)

Linear Predictive Cepstral coefficients are coefficients used in cepstral analysis and are produced from linear predictive coding. Because the LPCC approach provides a tract of the human vocal model, it is used to record emotional information [57]. However, the noise resistance of LPCC is low[58].

5.4 Perceptual Linear Predictive (PLP)

An all-pole model used in Perceptual Linear Predictive approaches to the auditory spectrum of speech. This process uses a bank of bark filters. The bark scale primarily captures the auditory system's subjective perception of loudness. It is well known that this approach performs calculations more slowly than MFCC[59].

5.5 Gammatone Frequency Cepstral Coefficients (GFCC)

Based on the gammatone filter bank, which represents the basilar membrane as a collection of overlapping bandpass filters, where the gammatone frequency cepstral coefficients were developed [60]. The GFCCs are generated by using an array of Gammatone filters to separate the signal of input speech into the domain of time frequency, followed by an adown sampling operation along the time duration[61].

5.6 Cochlear Filter Cepstral Coefficients (CFCC)

The Cochlear Filter Cepstral Coefficients (CFCC) are according to recently created audio transformation in addition to a collection of modules to replicate the cochlea's signal processing capabilities. To overcome the acoustic mismatch issue between the training and testing environments, the speaker recognition employment using the feature of CFCC. The CFCC features significantly outperform -PLP and MFCC under white noise[62].

5.7 Relative spectral processing [RASTA]

The appropriate data is extracted from the audio signal using the RASTA approach. The work's primary objective is to make speech recognition systems more durable in environments with additive noise and real-time reverberation[63], [64]. The technique is frequently employed for input signals governed by external noise or speech with noise disturbance. To perform better, the RASTA must be paired with PLP[65].

5.8 Convolutional Block Attention Module (CBAM)

The Convolutional Block Attention Module (CBAM) represents an attention module for convolutional neural networks. The CBAM is a compact, simple, and general model so that it can be incorporated with other model architectures, such as CNN and DNN models [66] .

6. SPEECH PROGRESSION THROUGH 1960s TO 2000s

The following is a summary of the areas in which automatic speaker identification technology has advanced during the last 50 years:

6.1 1960s and 1970s

The following summarizes the areas in which automatic speaker identification technology has advanced during the last 50 years. Ten years after automatic speech recognition, in the 1960s, the first attempts were made at the recognition of the speaker by applying correlation and a bank of filters. Pruzansky at Bell Labs[37], [67] was one of the first to research on the subject. One of the most critical issues in speaker recognition is intra-speaker variability of characteristics, which has been well studied[68]**.**

Text-independent methods: By averaging over a sufficient amount of time using statistical factors, different features were extracted to obtain speaker characteristics independent of the phonological environment. These consist of the following: spectrum and fundamental frequency histograms[69], instantaneous spectra covariance matrix[70], averaged auto-correlation[71], and spectral of long-term [72]. Approaches of text-dependent: The process of text-dependent was examined because the text-independent method's performance was constrained[69]–[71]. Comparing two exact text sound utterances in the same pronunciation settings can be done precisely and reliably using time-domain approaches when there is sufficient time alignment. As a result, text-dependent techniques perform far better than text-independent techniques.

Text-dependent approaches: Time-domain and text-dependent methods were also examined because the text-independent method's performance was constrained[69]–[71]. Comparing two utterances of the same text in similar phonetic settings can be done precisely and reliably using time-domain approaches when there is sufficient time alignment. As a result, text-dependent techniques perform far better than text-independent techniques.

6.2 1980s

Collection of VQ, can be effectively created by a series of speaker's vectors of training of short–term merits [73]. HMM as a parametric model was studied. An ergonomic HMM was suggested by Pritz [74]. A single-state HMM, currently known as the Gaussian mixture model (GMM), was suggested by Rose et al. [75] as a reliable parametric model.

6.3 1990s

In the 1990s, the study of enhancing adaptability grew as a major area of interest. The resistance against utterance fluctuations was the main point of comparison between the continuous or discrete ergonic HMM-based method and the VQ method, as reported by Matsui et al [76]. It was demonstrated that independent of the number of states, speaker recognition rates were highly associated with the overall number of states. Thus, GMM obtains nearly the same performance as the multiple-state ergodic HMM since employing information about transitions between states is ineffective for text-independent speaker recognition.

As an extension of speaker recognition technology, studies have emerged on extracting each person's speech intervals independently of conversation or meeting involving more than two individuals[77], [78]. Speaker recognition systems are becoming increasingly reliant on speaker segmentation and grouping approaches.

6.4 2000s

Features of High-Level, pronunciation, prosody, word idiolect, phone usage, and other high-level features have been effectively employed in text-independent speaker verification. Generally, high-level feature recognition algorithms use sound to generate a series of symbols, which are subsequently recognized based on their frequency and co-occurrence[34], [79]- [81].

Pronunciation, word idiolect, phone usage, and other high-level features have been effectively employed in the verification of speaker for text-independent. Generally, high-level feature recognition algorithms use sound to generate a series of symbols, which are subsequently recognized based on their frequency and co-occurrence[34][79], [80].

A range of adaptations of the model and compensating strategies were studied for speaker recognition approaches based on GMM. Two methods for GMM mean super-vector classifiers were proposed by McLaren et al.[82]. An unsupervised model adaptation strategy based on the posterior possibility that a test utterance corresponds to the model of the client was proposed by Petri et al. [83]- [85].

Integration of audio and visual characteristics: Audios-visual speaker verification systems, which use voice and image data together, have uttered a lot of interest. Lib movement is extremely employed as visual information. Enhancing system dependability is the audio-visual mix. Mixing two different information sources (audio-visual) able to be approached as a pattern classification problem or a classifier combination problem[86], [87], [88].

Adaptation of the model and compensating strategies were studied for GMM depending on techniques for speaker recognition[82]. An unsupervised model adaption strategy based on the target customer model was proposed by Peti et al. [89], [83].

Features of Maximum likelihood Linear Recognition (MLLR) to improve the robustness of voice recognition, and supervised and unsupervised HMM adaptation have both made extensive use of MLLR [90]. Promising experimental results were reported by Stolke [90], who suggested employing the adaptation matrix of MLRR as the speaker features [92]**.**

Table 1 Advancement of speaker recognition over the last decade

Year/Ref.	Method	Feature extraction	No. of speaker	Speaker language	Accuracy
2011-[93]	GMM-UBM	MFCC	100	Marathi	80% for noisy data
2012-[94]	VQ	LPCC	20	China-Mandarin	94.67%
2012-[95]	DNN	MFCC, LPC	280/ Mel and Femel	English	85%
2013-[96]	HMM	MFCC	140	English	93%
2014-[63]	HMM/GMM	RASTA-MFCC	AURORA databases	English	Error rate:3%-1%
2014-[97]	CWT	CFCC	5, 14, 21,23	English	Error 6.25%
2015-[98]	DNN	PLP	300	English	0.22%
2017-[99]	GMM	MFCC	267	Arabic	86%
2017[100]	DNN	MFCC	100	English	97.3%
2018-[101]	HMM-GMM	MFCC	10	Arabic-Amazigh	EER of 6%
2019-[102]	SVM	MFCC-LPC	12	Kurdish	EER of 6%
2019-[103]	TDNN	MFCC, PLP	591	English	EER OF 5.56
2019-[104]	DWT	Fuzzy Logic and NN	50	Arabic	NNs are better
2020-[105]	NNC	Far-Field Text-Dependent	340	Chinese, English	3.29% EER
2020-[106]	X-Vector	MFCC	180	VoxCeleb celebrities	2.7
2020-[107]	DNN	CBAM	1,000	VoxCeleb celebrities	EER 2.03%
2020-[108]	-	Different methods	7365	VoxCeleb diverse lang.	-
2021-[109]	GMM	GFCC	24	English	80%
2022-[110]	DNN	CBAM	Variable	Different language	EER 0.645
2023-[111]	NNC	MFCC	100	English	89.23%
2022-[47]	GMM+Genetic selection	MFCC	1160	English-Short duration	97.24

2022-[112]	Cochlear	MFCC Δ and ΔΔ	40+40	Short phrase of Malaysian &Bangladish	noise-robust performance
2023-[1123]	X-Vector	LPCC	1251	Chines	0.98%

7. CONCLUSION

Vocal recognition research becomes a significant academic research field in the middle of the twentieth century. The independent speaker recognition system was created by combining the concept of SR with the development of computers.

The concept of SR technology based on computers was advanced in the early years of computing. The spectrogram represented the first significant development in computers based on SR.

Many opinions, including modeling approaches and feature extraction techniques, have been covered in this work. It is clear that the MFCC frequently yields positive outcomes in some conditions and can be combined with other techniques to improve the performance of SR systems. The selection of the method depends on the type of problem. To use an appropriate method, the developer should consider numerous factors. However, the review found that the GMM method remains the best and most advanced approach for SR systems.

Today, business initiatives in speaker biometrics are becoming prevalent worldwide. After so many years of study, speaker recognition is only just approaching complete practical reliability (2% EER). Therefore, biometric systems now require the development of new technologies due to their many applications, especially when the necessity for individual identity emerges.

REFERENCES

[1] D. A. Reynolds, “Speaker identification and verification using Gaussian mixture speaker models,” Speech Commun., vol. 17, no. 1–2, pp. 91–108, 1995. DOI: 10.1016/0167-6393(95)00009.

[2] A. D. Yarmey, M. J. Yarmey, and L. Todd, “Frances McGehee (1912--2004): The first earwitness researcher,” Percept. Mot. Skills, vol. 106, no. 2, pp. 387–394, 2008. DOI:10.2466/pms.106.2.387-394

[3] F. McGehee, “The reliability of the identification of the human voice,” J. Gen. Psychol., vol. 17, no. 2, pp. 249–271, 1937. DOI: 10.1080/00221309.1937.9917999

[4] F. McGehee, “An Experimental Study Voice Recognition,” J. Gen. Psychol., vol. 31, pp. 53–65,1944. DOI:10.1080/00221309.1944.10545219

[5] R. Potter, G. Kopp, and H. Green, “Technical Aspects of Visual Speech,” Bell Labs, New York, 1947.

[6] L. G. Kersta, “Voiceprint identification,” J. Acoust. Soc. Am., vol. 34, no. 5, p. 725, 1962.

[7] K. P. Li, J. E. Dammann, and W. D. Chapman, “Experimental studies in speaker verification, using an adaptive system,” J. Acoust. Soc. Am., vol. 40, no. 5, pp. 966–978, 1966. DOI: 10.1121/1.1910221

[8] G. Fant, “Acoustic theory of speech production,” Natur Mag., 1960. Doi.org/10.1007/978-94-011-4657-9_3

[9] G. Cabadaug and Ö. Karal, “A Comparative Study of FFT Based Frequency Estimation Using Different Interpolation Techniques,” Al-Rafidain Eng. J., vol. 28, no. 2, pp. 86–93, 2023. DOI: 10.33899/rengj.2023.139624.1250

[10] J. W. Cooley and J. W. Tukey, “An algorithm for the machine computation of complex Fourier series”, vol. 19, Math. Comput., p. 73, 1965.

[11] B. P. Bogert, “The Quefrency Alanysis of the Time Series for Echos: Cepstrum Pseudo-Auto-Covariance, CrossCepstrum, and Saphe Cracking,” Math Comput., vol. 19, pp. 209–243, 1963. DOI: 10.4236/jcc.2014.22012

[12] A. V Oppenheim, “Speech analysis-synthesis system based on homomorphic filtering,” J. Acoust. Soc. Am., vol. 45, no. 2, pp. 458–465, 1969. DOI: 10.1121/1.1911395

[13] A. Oppenheim and R. Schafer, “Homomorphic analysis of Speech,” IEEE Transactions on Audio and Electroacoustics, vol. 16, no. 2, pp. 221–226, Jun. 1968. doi:10.1109/tau.1968.1161965. DOI: 10.1109/TAU.1968.1161965

[14] R. W. Schafer and L. R. Rabiner, “Digital Representation of Speech,” Invit. Pap. Proc. IEEE, vol. 63, p. 4, 1975. DOI: 10.1109/PROC.1975.9799

[15] B. S. Atal, “Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification,” J. Acoust. Soc. Am., vol. 55, no. 6, pp. 1304–1312, 1974. DOI: 10.1121/1.1914702

[16] B. S. Atal and S. L. Hanauer, “Speech analysis and synthesis by linear prediction of the speech wave,” J. Acoust. Soc. Am., vol. 50, no. 2B, pp. 637–655,1971. DOI: 10.1121/1.1912679

[17] B. S. Atal, “Automatic speaker recognition based on pitch contours,” J. Acoust. Soc. Am., vol. 52, no. 6B, pp. 1687–1697, 1972. DOI: 10.1121/1.1913303.

[18] B. S. Atal, “Automatic recognition of speakers from their voices,” Proc. IEEE, vol. 64, no. 4, pp. 460–475, 1976. DOI: 10.1109/PROC.1976.10155

[19] R. M. Hanifa, K. Isa, and S. Mohamad, “A review on speaker recognition: Technology and

challenges," Comput. \& Electr. Eng., vol. 90, p. 107005, 2021. DOI: 10.1016/j.compeleceng.2021.107005

[20] A. S. Ponraj and others, "Speech Recognition with Gender Identification and Speaker Diarization," in 2020 IEEE International Conference for Innovation in Technology (INOCON), 2020, pp. 1–4. DOI: 10.1109/INOCON50539.2020.9298241

[21] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," IEEE Trans. Audio. Speech. Lang. Processing, vol. 20, no. 2, pp. 356–370, 2012. DOI: 10.1109/TASL.2011.2125954

[22] S. Sujiya and E. Chandra, "A review on speaker recognition," Int J Eng Technol, vol. 9, no. 3, pp. 1592–1598, 2017, DOI:10.21817/ijet/2017/v9i3/170903513.

[23] N. K. Kaphungkui and A. B. Kandali, "Text Dependent Speaker Recognition with Back Propagation Neural Network," Int. J. Eng. Adv. Technol., vol. 8, no. 5, pp. 1431–1434, 2019. ISBN:0-7803-9313-9

[24] T. F. Zheng and L. Li, "Robustness-related issues in speaker recognition," vol. 2. Springer, 2017. DOI:10.1007/978-981-10-3238-7

[25] N. Singh, "Voice biometric: revolution in field of security," CSI Commun., vol. 43, no. 8, pp. 24–25, 2019. ISs N0970-647X

[26] A. M. Sharma, "Speaker recognition using machine learning techniques," M.S. thesis, Comp. Sci. Dept., San José State University, California, USA 2019. DOI: 10.31979/etd.fhhr-49pm

[27] P. K. Sharma et al., "Eminent method of voice identification by applying pitch, intensity and pulse," in AIP Conference Proceedings, 2022, vol. 2393, no. 1. DOI: 10.1063/5.0074174

[28] G. R. Doddington, "Speaker recognition Identifying people by their voices," Proc. IEEE, vol. 73, no. 11, pp. 1651–1664, 1985. DOI: 10.1109/PROC.1985.13345

[29] M. Jin and C. D. Yoo, "Speaker verification and identification," in Behavioral Biometrics for Human Identification: Intelligent Applications, IGI Global, 2010, pp. 264–289. DOI: 10.4018/978-1-60566-725-6.ch013

[30] N. Dehak, P. Dumouchel, and P. Kenny, "Modeling prosodic features with joint factor analysis for speaker verification," IEEE Trans. Audio. Speech. Lang. Processing, vol. 15, no. 7, pp. 2095–2103, 2007. DOI: 10.1109/TASL.2007.902758

[31] Y. Lukic, C. Vogt, O. Dürr, and T. Stadelmann, "Speaker identification and clustering using convolutional neural networks," in 2016 IEEE 26th international workshop on machine learning for signal processing (MLSP), 2016, pp. 1–6. DOI: 10.1109/MLSP.2016.7738816

[32] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," Digit. Signal Process., vol. 10, no. 1–3, pp. 19–41, 2000. DOI: 10.1006/dspr.1999.0361

[33] M. Kotti, E. Benetos, and C. Kotropoulos, "Computationally efficient and robust BIC-based speaker segmentation," IEEE Trans. Audio. Speech. Lang. Processing, vol. 16, no. 5, pp. 920–933,2008. DOI:10.1109/TASL.2008.925152

[34] G. Doddington, "Speaker recognition based on idiolectal differences between speakers," in Seventh European Conference on Speech Communication and Technology, 2001. (Eurospeech 2001), pp. 2521-2524, DOI 10.21437/Eurospeech.2001-417

[35] S. Mizuta and K. Nakajima, "A discriminative training method for continuous mixture density HMMs and its implementation to recognize noisy speech," J. Acoust. Soc. Japan, vol. 13, no. 6, pp. 389–393, 1992. DOI: 10.1250/ast.13.389

[36] N. Najkar, F. Razzazi, and H. Sameti, "A novel approach to HMM-based speech recognition systems using particle swarm optimization," Math. Comput. Model., vol. 52, no. 11–12, pp. 1910–1920, 2010. DOI: 10.1016/j.mcm.2010.03.041

[37] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE, vol. 77, no. 2, pp. 257–286, 1989. DOI: 10.1109/5.18626

[38] V. Vapnik, Estimation of dependences based on empirical data. Springer Science & Business Media, 2006. DOI: 10.1007/978-0-387-34239-9

[39] C. Cortes and V. Vapnik, "Support-vector networks, 1995," Mach. Learn., vol. 20, no. 3, p. 273, 1995. DOI: 10.1007/BF00994018

[40] L. Gbadamosi, "Voice Recognition System Using Template Matching," Int. J. Res. Comput. Sci., vol. 3, no. 5, p. 13, 2013. DOI: 10.7815/ijorcs. 35.2013.070.

[41] M. McLaren, Y. Lei, and L. Ferrer, "Advances in deep neural network approaches to speaker recognition," in 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2015, pp. 4814–4818. DOI: 10.1109/ICASSP.2015.7178885

[42] M. McLaren, L. Ferrer, and A. Lawson, "Exploring the role of phonetic bottleneck features for speaker and language recognition," in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5575–5579. DOI: 10.1109/ICASSP.2016.7472744.

[43] N. M. Almarshady, A. A. Alashban, and Y. A. Alotaibi, "Analysis and Investigation of Speaker Identification Problems Using Deep Learning Networks and the YOHO English Speech Dataset," Appl. Sci., vol. 13, no. 17, p. 9567, 2023. DOI: 10.3390/app13179567

[44] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," IEEE Trans. Acoust., vol. 37, no. 3, pp. 328–339, 1989. DOI:10.1016/b978-0-08-051584-7.50037-1

[45] D. A. Reynolds, Ph.D. dissertation, "A Gaussian mixture modeling approach to text-independent speaker identification". Georgia Institute of

Technology, 1992.

[46] T. THAT, “Automatic speaker recognition using Gaussian mixture speaker models,” Lincoln Lab. J., vol. 8, no. 2, 1995. Doi: 10.1121/1.2027823

[47] K. A. Kamiński and A. P. Dobrowolski, “Automatic Speaker Recognition System Based on Gaussian Mixture Models, Cepstral Analysis, and Genetic Selection of Distinctive Features,” Sensors, vol. 22, no. 23, p. 9370, 2022. DOI: 10.3390/s22239370

[48] Đ GROZDIĆ, S Jovičić, Z ŠARIĆ, I Subotić, “Comparison of GMM/UBM and i-vector based speaker recognition systems,” Sp 5th International Conference on Fundamental and Applied Aspects of Speech and Language Belgrade17-18 October, 2015 SPEECH Lang. 2015, p. 274, 2015.

[49] S. Mallat, A wavelet tour of signal processing. Elsevier, 1999. Academic Press 84 Theobald Road, London WCIX 8RR, UK. ISBN-13: 978-0-12-466606-1.

[50] P. Král, “Discrete Wavelet Transform for automatic speaker recognition,” in 2010 3rd International Congress on Image and Signal Processing, 2010, vol. 7, pp. 3514–3518. DOI: 10.1109/CISP.2010.5646691

[51] J. E. W. Koh et al., “Diagnosis of retinal health in digital fundus images using continuous wavelet transform (CWT) and entropies,” Comput. Biol. Med., vol. 84, pp. 89–97, 2017. DOI: 10.1016/j.compbiomed.2017.03.008

[52] R. Féraud and F. Clérot, “A methodology to explain neural network classification,” Neural networks, vol. 15, no. 2, pp. 237–246, 2002. DOI:10.1016/S0893-6080(01)00127-7

[53] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu, “Applications of support vector machine (SVM) learning in cancer genomics,” Cancer genomics & proteomics, vol. 15, no. 1, pp. 41–51, 2018. DOI:10.21873/cgp.20063

[54] L. Gong, S. Xie, Y. Zhang, Y. Xiong, X. Wang, and J. Li, “A Robust Feature Extraction Method for Sound Signals Based on Gabor and MFCC,” in 2022 6th International Conference on Communication and Information Systems (ICCIS), 2022, pp. 49–55. DOI: 10.1109/ICCIS56375.2022.9998146

[55] A. H. Abdulqader, S. A. Al-Haddad, S. Abdo, A. Abdulghani, and S. Natarajan, “Hybrid Feature Extraction MFCC and Feature Selection CNN for Speaker Identification Using CNN: A Comparative Study,” in 2022 2nd International Conference on Emerging Smart Technologies and Applications (eSmarTA), 2022, pp. 1–6. DOI: 10.1109/eSmarTA56775.2022.9935422

[56] J. Bradbury, “Linear predictive coding,” Mc G. Hill, 2000.

[57] K. T. Al-Sarayreh, R. E. Al-Qutaish, and B. M. Al-Kasasbeh, “Using the sound recognition techniques to reduce the electricity consumption in highways,” J. Am. Sci., vol. 5, no. 2, pp. 1–12, 2009. DOI:10.7537/marsjas050209.01

[58] M. M. El Choubassi, H. E. El Khoury, C. E. J. Alagha, J. A. Skaf, and M. A. Al-Alaoui, “Arabic speech recognition using recurrent neural networks,” in Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology (IEEE Cat. No. 03EX795), 2003, pp. 543–547. DOI: 10.1109/ISSPIT.2003.1341178

[59] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech,” J. Acoust. Soc. Am., vol. 87, no. 4, pp. 1738–1752, 1990. DOI: 0.1121/1.399423.

[60] N. Kanthi, “Speaker Identification based on GFCC using GMM,” Int. J. Innov. Res. Adv. Eng., vol. 1, no. 8, pp. 224–232, 2014. ISSN: 2277 128X.

[61] d Z. Arsalane, “Gammatone frequency cepstral coefficients for speaker identification over VoIP networks,” in 2016 International Conference on Information Technology for Organizations Development (IT4OD), 2016, pp. 1–5. DOI: 10.1109/IT4OD.2016.7479293

[62] Q. Li and Y. Huang, “An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions,” IEEE Trans. Audio. Speech. Lang. Processing, vol. 19, no. 6, pp. 1791–1801, 2010. DOI: 10.1109/TASL.2010.2101594

[63] H. Rahali, Z. Hajaiej, and N. Ellouze, “Robust Features for Impulsive Noisy Speech Recognition Using Relative Spectral Analysis,” Int. J. Electron. Commun. Eng., vol. 8, no. 9, pp. 1586–1591, 2014. DOI.org/10.5281/zenodo.1095933

[64] P. K. Kurzekar, R. R. Deshmukh, V. B. Waghmare, and P. P. Shrishrimal, “A comparative study of feature extraction techniques for speech recognition system,” Int. J. Innov. Res. Sci. Eng. Technol., vol. 3, no. 12, pp. 18006–18016, 2014.

[65] B. Singh, R. Kaur, N. Devgun, and R. Kaur, “The process of feature extraction in automatic speech recognition system for computer machine interaction with humans: a review,” Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 2, no. 2, pp. 1–7, 2012.

[66] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “Cbam: Convolutional block attention module,” in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19. DOI:/10.1007/978-3-030-01234-2_1

[67] S. Pruzansky and M. V Mathews, “Talker-recognition procedure based on analysis of variance,” J. Acoust. Soc. Am., vol. 36, no. 11, pp. 2041–2047, 1964. DOI: 10.1121/1.1919320

[68] W. Endress, W. Bambach, and G. Flosser, “Voice spectrograms as a function of age,” Voice Disguise Voice Imitation, JASA, vol. 49, no. 6, p. 2, 1971. DOI: 10.1121/1.1912589

[69] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. Acoust., vol. 27, no. 2, pp. 113–120, 1979. DOI: 10.1109/TASSP.1979.1163209

[70] K.-P. Li and G. W. Hughes, “Talker differences as they appear in correlation matrices of continuous speech spectra,” J. Acoust. Soc. Am.,

vol. 55, no. 4, pp. 833–837, 1974. DOI: 10.1121/1.1914608

[71] P. D. Bricker et al., “Statistical techniques for talker identification,” Bell Syst. Tech. J., vol. 50, no. 4, pp. 1427–1454, 1971. DOI:10.1002/j.1538-7305.1971.tb02561.x

[72] J. Markel, B. Oshika, and A. Gray, “Long-term feature averaging for speaker recognition,” IEEE Trans. Acoust., vol. 25, no. 4, pp. 330–337, August 1977 . DOI: 0.1109/TASSP.1977.1162961

[73] F. K. Soong, A. E. Rosenberg, B.-H. Juang, and L. R. Rabiner, “Report: A vector quantization approach to speaker recognition,” AT\&T Tech. J., vol. 66, no. 2, pp. 14–26, 1987. DOI: 10.1002/j.1538-7305.1987.tb00198.x

[74] A. Poritz, “Linear predictive hidden Markov models and the speech signal,” in ICASSP’82. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1982, vol. 7, pp. 1291–1294. DOI:10.1109/ICASSP.1982.1171633

[75] R. C. Rose and D. A. Reynolds, “Text independent speaker identification using automatic acoustic segmentation,” in International Conference on Acoustics, Speech, and Signal Processing, 1990, pp. 293–296. DOI: 10.1109/ICASSP.1990.115638

[76] S. Nakagawa and H. Suzuki, “A new speech recognition method based on VQ-distortion measure and HMM,” in 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1993, vol. 2, pp. 676–679. DOI: 10.1109/ICASSP.1993.319401

[77] H. Gish, M.-H. Siu, and J. R. Rohlicek, “Segregation of speakers for speech recognition and speaker identification.,” in icassp, 1991, vol. 91, pp. 873–876. DOI: 10.1109/ICASSP.1991.150477

[78] M.-H. Siu, G. Yu, and H. Gish, “An unsupervised, sequential learning algorithm for the segmentation of speech waveforms with multiple speakers,” in Acoustics, Speech, and Signal Processing, IEEE International Conference on, 1992, vol. 2, pp. 189–192. DOI Bookmark: 10.1109/ICASSP.1992.226088

[79] F. Bimbot et al., “A tutorial on text-independent speaker verification,” EURASIP J. Adv. Signal Process., vol. 2004, pp. 1–22, 2004. DOI: 10.1155/s1110865704310024

[80] S. Furui, “50 years of progress in speech and speaker recognition research,” ECTI Trans. Comput. Inf. Technol., vol. 1, no. 2, pp. 64–74, 2005. DOI:/10.37936/ecti-cit.200512.51834

[81] W. M. Campbell, J. R. Campbell, D. A. Reynolds, D. A. Jones, and T. R. Leek, “High-level speaker verification with support vector machines,” in 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004, vol. 1, pp. 1-73. DOI: 10.1109/ICASSP.2004.1325925

[82] M. McLaren, R. Vogt, B. Baker, and S. Sridharan, “A comparison of session variability compensation techniques for SVM-based speaker recognition,” in Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007), 2007, pp. 790–793.

[83] A. Preti, J.-F. Bonastre, D. Matrouf, F. Capman, and B. Ravera, “Confidence measure based unsupervised target model adaptation for speaker verification,” in Eighth Annual Conference of the International Speech Communication Association, 2007. DOI:10.21437

[84] X. Xie, X. Liu, H. Chen, and H. Wang, “Unsupervised Model-based speaker adaptation of end-to-end lattice-free MMI model for speech recognition,” in ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095083

[85] M. A. Al-Zakarya and Y. F. Al-Irhaim, “Unsupervised and Semi-Supervised Speech Recognition System: A Review,” AL-Rafidain J. Comput. Sci. Math., vol. 17, no. 1, pp. 34–42, 2023. DOI:10.33899/csmj.2023.179466

[86] M.-C. Cheung, M.-W. Mak, and S.-Y. Kung, “A two-level fusion approach to multimodal biometric verification,” in Proceedings. (ICASSP’05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005., 2005, vol. 5, pp. v--485. DOI: 10.1109/ICASSP.2005.1416346.

[87] 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, pp. 210–229, 2006. DOI: 10.1016/j.csl.2005.06.003

[88] M. A. Al-yoonus and S. A. Al-Kazzaz, “FPGA-SoC Based Object Tracking Algorithms: A Literature Review,” Al-Rafidain Eng. J., vol. 28, no. 2, pp. 284–295, 2023. DOI:10.33899/rengj.2023.138936.1243.

[89] P. Gimeno, D. Ribas, A. Ortega, A. Miguel, and E. Lleida, “Unsupervised adaptation of deep speech activity detection models to unseen domains,” Appl. Sci., vol. 12, no. 4, p. 1832, 2022. DOI:10.3390/app12041832

[90] S. J. Kuntz and J. B. Rawlings, “Maximum likelihood estimation of linear disturbance models for offset-free model predictive control,” in 2022 American Control Conference (ACC), 2022, pp. 3961–3966. DOI: 10.23919/ACC53348.2022.9867344

[91] A. S. L. Ferrer and S. K. E. S. A. Venkataraman, “MLLR Transforms as Features in Speaker Recognition,” 2006. Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, USA. ID: 5180980

[92] M. Wolf, “Channel selection and reverberation-robust automatic speech recognition,” PhD Thesis, Dept. of Signal Theory and Communication, Universitat Polit`ecnica de Catalunya, Barcelona, Spain, 2013. DOI: 10.5821/dissertation-2117-95257

[93] P. Krishnamoorthy, H. S. Jayanna, and S. R. M. Prasanna, “Speaker recognition under limited data condition by noise addition,” Expert Syst.

Appl., vol. 38, no. 10, pp. 13487–13490, 2011. DOI: 10.1016/j.eswa.2011.04.069

[94] L. Zhu and Q. Yang, “Speaker Recognition System Based on weighted feature parameter,” Phys. Procedia, vol. 25, pp. 1515–1522, 2012. DOI: 10.1016/j.phpro.2012.03.270

[95] B. P. Das and R. Parekh, “Recognition of isolated words using features based on LPC, MFCC, ZCR and STE, with neural network classifiers,” International Journal of Modern Engineering Research (IJMER) www.ijmer.com Vol.2, Issue.3, May-June 2012, pp. 854-858 ISSN: 2249-6645

[96] S. Bhardwaj, S. Srivastava, M. Hanmandlu, and J. R. P. Gupta, “GFM-based methods for speaker identification,” IEEE Trans. Cybern., vol. 43, no. 3, pp. 1047–1058, 2013. DOI: 10.1109/TSMCB.2012.2223461

[97] N. Singh, N. Bhendawade, and H. A. Patil, “Novel cochlear filter based cepstral coefficients for classification of unvoiced fricatives,” Int. J. Nat. Lang. Comput, vol. 3, no. 4, pp. 21–40, 2014. DOI : 10.5121/ijnlc.2014.3402

[98] Y. Liu, Y. Qian, N. Chen, T. Fu, Y. Zhang, and K. Yu, “Deep feature for text-dependent speaker verification,” Speech Commun., vol. 73, pp. 1–13, 2015. DOI: 10.1016/j.specom.2015.07.003

[99] M. Alsulaiman, A. Mahmood, and G. Muhammad, “Speaker recognition based on Arabic phonemes,” Speech Commun., vol. 86, DOI:10.1016/j.specom.2016.11.004.

[100] W. Jiang, P. Liu, and F. Wen, “Speech magnitude spectrum reconstruction from MFCCs using deep neural network,” Chinese J. Electron., vol. 27, no. 2, pp. 393–398, 2018. DOI: 10.1049/cje.2017.09.018

[101] N. Singh, A. Agrawal, and R. Khan, “The development of speaker recognition technology,” IJARET., Vol. 9, Issue: 3, pp. 8–16, May – June 2018.

[102] S. M. Omer, J. A. Qadir, and Z. K. Abdul, “Uttered Kurdish digit recognition system,” J. Univ. Raparin, vol. 6, no. 2, pp. 78–85, 2019. DOI: 10.26750/paper

[103] C.-P. Chen, S.-Y. Zhang, C.-T. Yeh, J.-C. Wang, T. Wang, and C.-L. Huang, “Speaker characterization using tdnn-lstm based speaker embedding,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 6211–6215.DOI: 10.1109/ICASSP.2019.8683185.

[104] L. Eljawad et al., “Arabic voice recognition using fuzzy logic and neural network,” Eljawad, L., Aljamaeen, R., Alsmadi, MK, Al-Marashdeh, I., Abouelmagd, H., Alsmadi, S., Haddad, F., Alkhasawneh, RA, Alzughoul, M. & Alazzam, MB, pp. 651–662, 2019. DOI: org/10.1063/5.0094741

[105] X. Qin, H. Bu, and M. Li, “Hi-mia: A far-field text-dependent speaker verification database and the baselines,” in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 7609–7613. DOI: 10.1109/ICASSP40776.2020.9054423

[106] Q.-B. Hong, C.-H. Wu, H.-M. Wang, and C.-L. Huang, “Statistics pooling time delay neural network based on x-vector for speaker verification,” in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6849–6853.DOI: 10.1109/ICASSP40776.2020.9054350

[107] S. Yadav and A. Rai, “Frequency and temporal convolutional attention for text-independent speaker recognition,” in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6794–6798. DOI: 10.1109/ICASSP40776.9054440.

[108] V. Vestman, T. Kinnunen, R. G. Hautamäki, and M. Sahidullah, “Voice mimicry attacks assisted by automatic speaker verification,” Comput. Speech \& Lang., vol. 59, pp. 36–54, 2020. DOI: 10.1016/j.csl.2019.05.005

[109] U. Kumaran, S. Radha Rammohan, S. M. Nagarajan, and A. Prathik, “Fusion of mel and gammatone frequency cepstral coefficients for speech emotion recognition using deep C-RNN,” Int. J. Speech Technol., vol. 24, pp. 303–314, 2021. DOI: 10.1007/s10772-020-09792-x

[110] X. Qin, N. Li, C. Weng, D. Su, and M. Li, “Simple attention module based speaker verification with iterative noisy label detection,” in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 6722–6726. DOI: 10.1109/ICASSP43922.2022.9746294

[111] G. Costantini, V. Cesarini, and E. Brenna, “High-Level CNN and Machine Learning Methods for Speaker Recognition,” Sensors, vol. 23, no. 7, p. 3461, 2023. DOI: 10.3390/s23073461

[112] M. A. Islam, Y. Xu, T. Monk, S. Afshar, and A. van Schaik, “Noise-robust text-dependent speaker identification using cochlear models,” J. Acoust. Soc. Am., vol. 151, no. 1, pp. 500–516, 2022. DOI: 10.1121/10.0009314

[113] Y. Zhang and L. Liu, “Multi-task learning for X-vector based speaker recognition,” Int. J. Speech Technol., pp. 1–7, 2023. DOI: 10.1007/s10772-023-10058-5

طرق وتقنيات تمييز التحدث: بحث مراجعة

عبد العليم عبد الفتاح رشيد
alem12@uomosul.edu.iq

محمد طارق ياسين
mtyaseen@uomosul.edu.iq

مروان احمد عبد الحميد
marwanhajali@uomosul.edu.iq

قسم الهندسة الكهربائية، كلية الهندسة، جامعة الموصل، موصل، العراق

تاريخ الاستلام: 12 نوفمبر 2023 استلم بصيغته المنقحة: 24 ديسمبر 2023 تاريخ القبول: 6 يناير 2024

الملخص

يعرف نظام التحقق من الهوية وتحديد الهوية بناء على الخصائص الصوتية المميزة للشخص باسم التعرف على المتحدث. تلقي هذه الورقة على تطور انظمة التعرف على المتحدثين من الايام الاولى لأجهزة الكمبيوترالى احدث الابتكارات. يمثل الصوت مقياس السلوك الحيوي الذي ينقل تفاصيل سمات الشخص, بدءا من عمر المتحدث والجنس والعرق. يركز مجال التعرف على المتحدث على تحديد الافراد من خلال اصواتهم. كان التعرف على التحدثين بشكل كاف موضوعا للبحث على مدار العقود الثمانية الماضية, جعلت التطبيقات مثل انترنت الاشياء, والمنازل الذكية, والادوات الذكية استخداما عصريا في العصر الحديث. نحن نقدم نظرة عامة سريعة على مجال التعرف على المتحدث في هذا العمل من خلال تحديد منهجية النمذجة واستراتيجيات استخراج الميزات المختلفة باستخدام مجموعة متنوعة من اللغات. الهدف من ادبيات التعرف على المتحدث هو تعزيز المعرفة الاكاديمية للتعرف على المتحدث.

الكلمات الداله

تمييز التحدث, البصمة، MFCC، GMM، استخلاص الميزات