DOI: 10.24996/ijs.2021.62.9.38

ISSN: 0067-2904

Analysis of Methods and Techniques Used for Speaker Identification, Recognition, and Verification: A Study on Quarter-Century Research Outcomes

Thabit Sultan Mohammed[1*], Karim M. Aljebory[1], Mohammed Aref Abdul Rasheed[2]
Muzhir Shaban Al-Ani[3], Ali Makki Sagheer[1]

[1*]Computer Technical Engineering Dept., Al-Qalam University College, Kirkuk, Iraq.
[2]MIS Dept., College of Commerce and Business, Dhofar University, Oman.
[3]Department of IT, University of Human Development, Sulaimani, KRG, Iraq.

Abstract

The theories and applications of speaker identification, recognition, and verification are among the well-established fields. Many publications and advances in the relevant products are still emerging. In this paper, research-related publications of the past 25 years (from 1996 to 2020) were studied and analysed. Our main focus was on speaker identification, speaker recognition, and speaker verification. The study was carried out using the Science Direct databases. Several references, such as review articles, research articles, encyclopaedia, book chapters, conference abstracts, and others, were categorized and investigated. Summary of these kinds of literature is presented in this paper, together with statistical analyses to represent the publications and their categories over the mentioned period. Important information, including the dataset used, the size of the data adopted, the implemented methods, and the accuracy of the obtained results in the analysed research, are extracted from the explored publications and tabulated. The results show that the sum of published research articles is outnumbering other categories of publications. The number of researches in speech and speaker identification, recognition, and verification shows an increasing trend. Based on the normalized comparative factors of research publications, we found that many of them reached a high level of accuracy in their findings; hence the significantly superior techniques were derived and discussed for future researches. This survey paper would be beneficial for all those who wish to enhance their researches in the area of voice identification, recognition, and verification.

Keywords: Speaker identification, Speaker recognition, Speaker verification, Speech processing.

تحليل الأساليب والتقنيات المستخدمة في تحديد المتحدث والتعرف عليه والتحقق منه: دراسة حول نتائج البحث خلال ربع قرن

ثابت سلطان محمد [1] [*], كريم محمد الجبوري[1] ,محمد عارف عبد الرشيد [2] , مزهر شعبان العاني [3] , علي مكي صغير[1]

[1] * قسم هندسة تقنيات الحاسوب ، كلية القلم الجامعة ، كركوك ، العراق.

[2] قسم نظم المعلومات الإدارية ، كلية التجارة والأعمال ، جامعة ظفار ، سلطنة عمان.

*Email: thabit.sultan@alqalam.edu.iq

[3] قسم تقنية المعلومات ، جامعة التنمية البشرية ، السليمانية ، إقليم كردستان العراق.

الخلاصة

تعد نظريات وتطبيقات تحديد هوية المتحدث، والتعرف عليه، والتحقق منه من بين المجالات الراسخة ، ولا يزال هناك الكثير من البحوث المنشورة ، والمزيد من النتاجات العلمية ذات الصلة آخذة في الظهور. في هذا البحث ، تمت دراسة وتحليل الادبيات المنشورة والمتعلقة بالبحوث على مدار الـ 25 عامًا الماضية من 1996 إلى 2020. تركز هذه الورقة البحثية بشكل رئيسي على تحديد هوية المتحدثين ، وعلى التعرف على المتحدث ، والتحقق من المتحدث. أجريت الدراسة باستخدام قواعد بيانات مؤسسة Science Direct. حيث تم تصنيف العديد من المراجع كمقالات المراجعة والتقييم النقدي، ومقالات بحثية، وموسوعة وفصول كتب وملخصات مؤتمرات وغيرها. احتوت الورقة على ملخصات استعرضت هذه الأنواع من الأدبيات، جنبًا إلى جنب مع التحليلات الإحصائية لتمثيل المنشورات وفئاتها خلال الفترة المذكورة. تم استخلاص المعلومات المهمة ، بما في ذلك مجموعة البيانات المستخدمة ، وحجم البيانات المعتمدة ، والطرق المنفذة ، ودقة النتائج التي تم الحصول عليها في البحوث التي تم تحليلها ، ومن المنشورات التي تم استكشافها وجدولتها. توقد اضهرت النتائج أن مجموع المقالات البحثية المنشورة يفوق عدد فئات المنشورات الأخرى، وكذلك أن عدد الأبحاث الخاصة بموضوع التعرف على الصوت والتعرف على المتكلم والتحقق منه اتجاهًا متزايدًا. استنادًا إلى العوامل المقارنة المعيارية للمنشورات البحثية ، ووجد أن العديد منها قد وصل إلى مستوى عالٍ من الدقة في نتائجها ؛ ومن ثم تم اشتقاق التقنيات الأفضل بشكل ملحوظ ومناقشتها لأبحاث مستقبلية. ستكون ورقة الاستطلاع هذه مفيدة لجميع أولئك الذين يرغبون في تعزيز أبحاثهم في مجال التعرف على الصوت والتعرف على المتحدث والتحقق منه.

1. Introduction

The secrecy and protection of information of every organization is utmost important for sustainability and continuous development purposes. The methods of data and information protection have evolved greatly as the traditional methods have been dispensed with the introduction of modern methods based on biometric characteristics. The biometric characteristics are distinct from one to other human being. They can be broadly divided up into physiological characteristics and may include fingerprint recognition, face recognition, ear form, DNA, and behavioural characteristics. Furthermore, biometric characteristics may also include speaker recognition, signature recognition, voice pattern, and gait recognition. They can be obtained and measured from a biometric sample for the purpose of biometric identification. Voice and speaker prints are important features related to behavioural characteristics for recognition, identification, and verifications of speakers. Many methods, algorithms, approaches, and datasets are used to identify speakers with accuracy, depending on many factors. Noise removal process is always applied at the beginning of the identification process to decrease the effect of noise and to gain better accuracy in the identification results.

A general speaker identification system is basically composed of two phases; training phase, and testing phase. These phases include several steps, as illustrated in Figure-1:

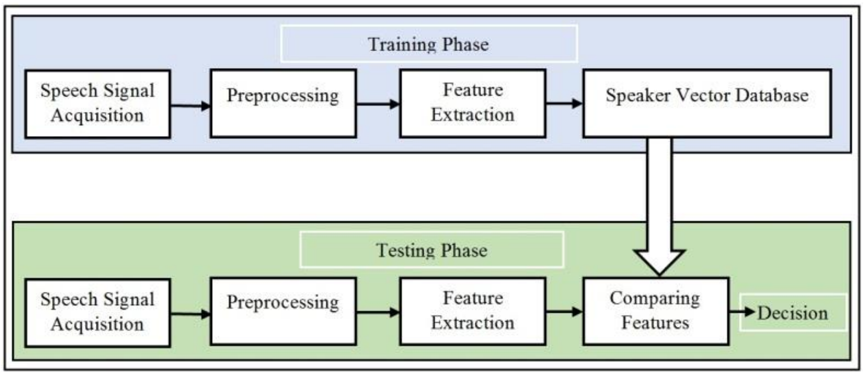


Figure -1 A general speaker identification system

- **Speech Signal Acquisition**: In this step, the speech voice is received by a microphone sensor and converted into an electrical signal. Then, the electrical signal is converted into a digital signal or data that is passed to the processing step.
- **Pre-processing**: The digital signal is divided up into samples and then resized. Moreover, a noise removal filter may also be used in this step.
- **Feature Extraction**: In this step, many methods are applied to extract features from the voice signal.
- **Speaker Vector Database**: In this step, feature vectors are saved in the database.
- **Comparing Features**: The new feature vector is compared with the database to select the similar matching vector, which indicates that a certain speaker is identified.

In this paper, three key words, namely speaker identification, speaker recognition, and speaker verification, are considered. The research works published in the Science Direct database were studied. Science Direct offers access to a large database of scientific and medical research with over 12 million items of content from 3,500 academic journals and 34,000 e-books. The related and required information were tracked down from various publications and analysed accordingly. As part of the survey of these publications, an extensive literature review is presented in section 2, while three search methods based on the adopted keywords together with analysis of results are presented in section 3. In subsections of section 3, graphs are drawn to illustrate a visualized comparison between the number of published items in terms of review articles, research articles, encyclopaedia, book chapters, conference papers, and others. A comparison for the published research articles within the key words' speaker identification, speaker recognition, and speaker verification, is presented in section 4. In section 5, tabular analysis of the searched published and examined works is presented. Tables are provided to illustrate a summarized analysis based on the references, authors, year of publication, the dataset used in the research and its size, the adopted method, and finally the accuracy of the obtained results. The analysis techniques adopted in the paper and the obtained results are discussed in section 6. Finally, in section 7, conclusions to our survey are presented.

2. Literature Review

In order to situate the work presented in our study within the body of the relevant literature, and to ensure that a useful context is provided, an overview of previously published works on speaker identification, recognition, and verification will be presented in this section. A focus is given to the research work that is published in the Science Direct database. The organization of the section depends on the years of publications, such that the most recent works are presented first.

2.1 Published works in the year 2020

Vestman *et al.* [1] have used the Automatic Speaker Verification (ASV) technology. The study was carried out on voice mimicry data. The comparison was performed between the voices of potential target speaker and hired attackers by adopting the x-vectors in the attacked side, while on the other side, i-vector technology was adopted. A similar ranking was studied that transfers from attackers to attacks by mimicking or imitating within the ASV system. Further investigation was conducted to track improvements in mimicking and the variation of properties in the voices of attackers during mimicking. Among their principal findings is that imitating does not show a progress in attacks, when the normal voices of attackers and targets are comparable to each other. Additionally, untrained impersonators do not posture a high threat towards ASV systems. A potential threat may exist in the case that an ASV system is utilized to assault other ASV systems.

Nanxin *et al*. [2] investigated some issues relevant to the reliability of automatic speaker recognition systems. Their research addressed the case of pairs of speakers that may be identical to each other in terms of ASV. They proposed a framework that allows predicting the safety of ASV technology in particular, where they analysed the performance of two ASV systems based on i-vector and x-vector speakers embedding. A number of 1000 target speakers were considered to generate up to 100,000 virtual impostor samples. They obtained the reasonable agreement between the false alarm rates of the generated samples and the empirical false alarm rates.

2.2 Published works in the year 2019

The speakers who need to pass on spoken messages over separation or above ecological commotion need to increase the strength of their voice by shouting. In the case of shouting, the normal speech, and the speaker identification system, may result in significantly low recognition performance.

Joinen *et al*. [3] addressed this issue, where they proposed two compensation methods to deal with the possible mismatch that a speaker recognition system may encounter for the abovementioned performance of the speaker identification system. Their proposed technique was demonstrated in the feature extraction stage, where they made changes to the spectral envelope’s high peak voices and brought it near to the normal voice. They illustrated a significantly enhanced rate in speaker identification.

Finding the gender of people from their voice for forensic purposes, based on specific information extracted from speech, is an issue that was investigated by many researchers. Under the noiseless and noisy conditions Kenai *et al*. [4] adopted various feature extraction techniques, and their system, which is referred to as Forensic Gender Speaker Recognition (FGSR), was evaluated in terms of Equal Proportion Probability (EPP) with sets of people composed of both genders. Performance evaluation results that were presented are more encouraging in noiseless conditions compared to noisy conditions.

A scoring method that is based on distance calculation for degree of similarity in text-independent speaker verification was presented by Hourri and Kharroubi [5]. Their approach stands as an alternative to stochastic models for text-independent speaker verification, which is comparatively expensive. A similarity measurement method was proposed using the speaker’s feature vectors (MFCC), in order to preserve and take advantage of the speaker’s specific features. Few experiments were conducted on open-source speaker recognition systems and demonstrated better performance compared to some well-known approaches.

A Dual Speaker Gesture (DuG) recognition system was presented by Ai *et al.* [6]. For this scheme, the sensing environment was used with two speakers with microphones. The purpose behind this system was to care for the environmental noise, nearby human motions, used to control robots. Such noise and human motion affects the accuracy of recognizing gestures and weaken the effectiveness of the interaction mode. The system is also having a fusion framework to combine a priori empirical model with a Hidden Markov Model [HMM) to enhance classification accuracy and improve user adaptability.

Speech features that are relevant to the physical framework of the speaker, together with some learned habits of speaking, are factors on which most speaker recognition systems operate. Sabatier *et al*. considered the case of identical twins of same gender having similar features obviously. They reported that only little research was carried out to measure the impact of the claim that identical or matching twins could lessen the capability of speaker recognition systems. In their work, 167 pairs if those indistinguishable twins were studied, and verification experiments using read and conversational speech samples were collected from those twins [7].

In various applications in which verification processes are required, like mobile devices and bank transactions, the biometric recognition has become familiar and widespread. The automatic speaker verification (ASV) allows individuals to verify their identity to an application by comparing live collected speech sampled data with reference information stored on the application's server. With all this in existence, preserving privacy in ASV is of great importance because biometric data, such as fingerprints or speech, may be used to collect a lot of sensitive information about the data subject. Treiber *et al*. addressed this issue and proposed architecture to enable privacy-preserving ASV in the encrypted domain. The authors stated that their proposed architecture is performing better than a previous scheme, namely the homomorphic encryption (HE), where the usage of the latter encryption comes with a rather heavy overhead, leading to a slow verification process [8].

2.3 Published works in the year 2018

George *et al*. used the Cosine Distance Feature (CDF) technique for extracting speech features to propose a distance-based representation. They achieved this by encoding the cosine distance between i-vectors of the utterances belonging to target speaker and reference speakers. They used CDF with a Support Vector Machine (SVM) classifier (CDF-SVM), and found that the reference speakers are more important to discriminate between speakers who are highly close in acoustic similarity to the target speaker. In addition to that, the speaker specific CDF showed that the acoustic similarities between the target and reference speakers are best captured using an intersection kernel SVM where each target speaker has specific subset of most acoustically similar reference speakers [9].

Vestman *et al*. presented a paper in which they addressed the problem of incompatible speaking styles made by the speaker. They focused on whispering being among the ways of hiding one's speaker identity. Although a whispered speech is normally clear, it is of low-intensity and therefore disposed to distortions. Results of the experiments applied in this research indicated that the testing result of normal-whisper mismatched conditions improves speaker recognition performance by 7–10% over the standard MFCC features in relative terms using the FDLP-TVLP features [10].

Implementation of speaker recognition for speech samples related to Hindi speakers was presented by Maurya *et al*. [11]. They used Mel frequency cepstral coefficient–vector quantization (MFCC-VQ) for the text dependent phase, while Mel frequency cepstral Coefficient-Gaussian mixture model (MFCC-GMM) was used for the text independent phase. In this research, more than 270 trails for both genders were tested, and the accuracy was found to be higher in the case of text dependent recognition.

Athulya and Sathidevi stated that when some investigations require speech biometrics as evidence, they are usually found to be highly distorted [12]. In their research, they considered a noticeable distortion introduced by the speech codec, where some of the speaker-specific features are either removed or distorted, causing reduction in the speaker verification accuracy. The paper quantified the effect of distortion on frequently used speaker-specific features, such as Mel Frequency Cepstral Coefficients (MFCC). The paper adopted the feature which is least affected by the codec of Power Normalized Cepstral Coefficients (PNCC) and proposed a slight modification as an improvement to the verification accuracy. The modified PNCCs (MPNCC) helped in reducing error rate.

For the purpose of speech recognition, a low dimensional speaker and channel dependent space was defined by Ibrahim and Ramli [13]. They used a simple i-vector factor analysis for the defined space, which was referred to as total variability space, since it models both speaker and channel variability. A database for an identification system that identifies frog's sound was used, with parameters of the system are initially tuned with some value. The effect of the tuned parameter was assessed and the computation time was recorded.

2.4 Published works in the year 2017

The SVM, Dynamic Time Warping (DTW), and Gaussian Mixture Model (GMM) techniques were used by Ding and Shin in [14] for verification, identification and recognition of speakers to develop a Kinect microphone array based to control humanoid robot via speech and speaker recognition process. With this proposed scheme, authenticated users can control a humanoid robot through voice commands. The user legitimacy has to be first verified by the robot, together with the command validity, before executing a voice command.

The task of speaker verification is undoubtedly difficult in deliberate situations in case of speakers purposely mask their own identity. Hautamäki *et al.* [15] have presented a research approaching voice mask or disguise from the view point of acoustical and perceptual analysis using a sample of 60 native male and female Finnish speakers producing utterances in normal, intended young and intended old voice modes. Among their investigations in this research is to study the effect of disguise as a relative change in fundamental frequency (F0) and formant frequencies (F1 to F4) from modal to disguised utterances. They also considered affecting factors from the listener's side. The research results indicated that what can be categorized as easy or difficult by an ASV system may also be categorized similarly for the average listener.

The investigation on the effect of Arabic phonemes on how speaker recognition systems are performing was studied by Alsulaiman *et al*. [16]. They revealed that some Arabic phonemes are strongly affecting the recognition rate of such systems. The researcher's findings showed that rates of recognition for Arabic vowels were all above 80%, while for the consonants, the rate varied from very low (14%) to very high (94%). To build a higher performance speaker recognition system, the recommendation of this research is to segment the most effective phonemes and use the research findings.

As a speaking style, the whispered speech is featured by its reduced possibly for proper recognition. It contains considerable information about the intended message and speaker identity and his/her gender that can lead to accepted recognition and hence verification. In this regard, Sarria-Paj and Falk [17], reported two limitations in relevant recognition systems. Firstly, the use of conventional features (e.g. mel-frequency cepstral coefficients, MFCCs) does not deliver adequate speaker judgement between whispered and normally-phonated speech. Secondly, training on both types of speech may enhance the whispered speech performance, but it will be on the expense of normal speech. Their research objective was to deal with the two stated limitations by recommending three new features, which when fused at the score level, gave results that are considered reliable for both types of speech, normal and whispered.

Ghoniem, and Shaalan [18] presented an Arabic text-independent speaker verification system. In this research, new speech features denoted as Wavelet Packet Four-Directional Features (WPFDF) were proposed for speaker characterization. Furthermore, a Fuzzy Hidden Markov Model (FHMM) was also proposed, aiming to enhance the speaker verification. Researchers here stated that the kernel fuzzy c-means (KFCM) is extended to calculate fuzzy memberships of HMMs training samples, leading to a reduction in information loss as well as a noticeable increase in the recognition rate.

A recognition system was presented by Paulose *et al*. [19], in which both spectro-temporal features and voice-source features are used for implementation. In this research, it was stated that the accuracy of recognition systems relies on the methods used in extracting features from the speech signal, modelling methods, classifiers used in speaker identification, as well as the amount of data available for training and testing. For the system proposed in this paper, two different classifiers were used and the accuracy rates were compared.

2.5 Published works in the year 2016

Wang *et al.* stated that degradation in speaker verification performance is found when the input speech is tested over a long period of time and at different sessions [20]. In order to overcome the problem, they presented speech database, which is especially collected to show how speaker's performance may vary in the long-term. They further explored and examined the issues in the frequency domain by underlining higher discrimination for speaker-specific information and lesser sensitivity to information that are time-related and session-specific. F-ratio was used as a criterion to find the figure of merit to judge the collected information and to determine a compromise between them. The F-ratio relates to the pre-filtering frequency warping and the post-filtering filter-bank outputs weighting. The proposed approach, which was claimed to perform well, was also tested by the use of the NIST SRE 2008 database.

Combination of the human senses of sight and hearing can be used to achieve what is referred to as situational awareness. Microphones for audio and cameras for video, in cooperation, were fused and integrated by D'Arca *et al.* [21] at semantic abstractions using different levels. The system presented in this research is to detect and follow a speaker who has a relative freedom to move within an area that is larger than a round table. Among the research findings is that the overall multimodal follower, which is tracking the speaker result in better reliability than single modality systems, and the advanced performance given by the audio–video integration and fusion, showed better tracking precision and speaker recognition.

A feature-level and score-level fusion approach was presented by Li *et al*. in [22]. For both text independent and text dependent speaker verification, the researchers combined acoustic and estimated articulatory information. In this study, the improvement of speaker verification performance by merging dynamic articulatory information with the conventional acoustic features was also presented. Because measuring of articulatory data is relatively having little feasibility in many real world applications, researchers were pushed to experiment with estimated articulatory features obtained through acoustic-to-articulatory inversion. Feature-level and score-level fusion methods were explored and the overall system performance demonstrated significant enhancement, even with estimated articulatory features. X-ray Microbeam database and the RSR 2015 database were used during experimentation.

Investigation on measures of speech quality that are related to the Speaker Verification (SV) performance was the motivation behind the research presented by Villalba *et al*. [23]. In that research, measures like modulation index, signal-to-noise ratio (SNR), number of speech frames, as well as shimmer, jitter, JFA, and probabilistic linear discriminant analysis models, were analysed. Furthermore, a measure based on the vector Taylor series (VTS) was proposed, while Bayesian networks were used to combine these measures and produce a probabilistic reliability measure. Bayesian network was trained on NIST SRE08 distorted with noise and evaluated on a distorted version of SRE10.

Rao and Mak were inspired by taking advantages of empirical kernel maps and incorporating them into a more advanced kernel machine called relevance vector machines (RVMs) [24]. Extensive analyses on the behaviours of RVMs were reported in this paper, together with providing insight into the properties of RVMs and their applications in i-vector/PLDA speaker verification. Their results stated that PLDA–RVM is much sparser than PLDA–SVM.

Since the variations in samples are among the main complications associated with speaker recognition, an update for both online and offline features together with model update techniques were considered by Anzar *et al.* [25]. They adapted the Vector Quantization (VQ) approach for feature update, while Gaussian Mixture Model (QMM) approach was considered for model updating. Anzar's *et al.* methods, which were updating the feature automatically in accordance with the biometric sample variations over time, improved the recognition accuracy and reduced the classification errors for voice recognition systems. The templates in the presented approach were continually adapted based on semi-supervised learning strategies.

2.6 Published works in the year 2015

The cases of mismatch conditions occurring during the performance of automatic speaker recognition were considered by Chougule and Chavan [26]. They proposed a spectral feature set, called NDSF (Normalized Dynamic Spectral Features), with a magnitude spectral subtraction being performed on spectral features for compensation against additive noise. They further applied modifications using time-difference approach followed by Gaussianization Non-linearity to compensate the effects of channel mismatch and handset transducers. The performance of the proposed feature set was compared with conventional cepstral features, like mel-frequency cepstral coefficients. From the studies presented in the research, which were performed on two databases, it was observed that the spectral domain dynamic features reduced the additive noise and channel effects caused by sensor mismatch and hence enhanced the robustness.

Approaches of extracting and using features from deep learning models for text-dependent speaker verification were presented by Liu *et al.* [27]. The motivation of their research was that deep learning has not been thoroughly explored and accepted for speaker verification as for speech recognition. In their approach, it was proposed that outputs from hidden layer of various deep models are employed as deep features for text-dependent speaker verification. Four types of deep models were investigated. The proposed approaches were evaluated on the RSR2015 data corpus. The experiments showed significant performance improvements compared to the traditional baselines.

The performance of three modern speaker verification systems and non-expert human listeners in the presence of voice mimicry were compared by Hautamäki *et al.* [28]. The goal of the research was to gain an understanding on how weak speaker verification systems are to mimicry attacks and to compare them to the performance of human listeners. The research, which adopted material in Finnish language for the study, found that the mimicry attack decreased slightly for Gaussian Mixture Model-Universal Background Model (GMM-UBM), while for i-vector systems, the Equal Error Rate (EER) increased. The performance of the human listening shows that mimicked speech increases the difficulty of the speaker verification task. It was also found difficult to recognize persons who are intentionally concealing their identity.

Haris and Sinha [29] explored the use of low-complexity data-independent estimates for reducing the dimensionality of GMM super vectors in context of speaker verification (SV). They adopted the NIST 2012 SRE task using a state-of-the-art PLDA based SV system and used sparse random matrix for driving their estimates. They further explored decimation and used them as speaker representations. They also proposed a speaker verification system that exploits the diversity among the representations obtained by using different offsets in the decimation of super vector.

In speech acquisition, it is often possible to practice clipping as a mean of storage utilization due to the limited numerical range or the non-linear compensation of recording devices. Clipping is unavoidably changing the spectrum of speech signals, causing partial distortion for the speaker information contained in the signal. Bie *et al.* investigated the impact of signal clipping on speaker recognition and proposed approaches for both clipping detection and signal reconstruction based on deep neural networks (DNNs) [30]. Results of this research reported that clipping affects the performance of speaker recognition, but the impact is slightly marginal if the clipping rate does not exceed 80%.

A common practiced operation to change people's voice and to conceal their identity is voice transformation. Despite the fact that this practice may present threats to security, few efforts have been reported on the recognition of hidden speakers from such disguised voices. Wang *et al.* (31) proposed countermeasures to erase the disguise effects and verify the speaker's identity from voice transformation disguised voices. The reported results of this proposed system stated that when countermeasures were adopted, the verification performances showed significant improvement with Equal Error Rate (EER) lowered to 3%–4%.

Despite the expanding motion to develop spoofing countermeasures for automatic speaker verification, the problem is still requiring more efforts towards effective solutions, and biometric systems remain vulnerable to spoofing. Wu *et al.* [32] presented a survey of relevant literature and identified the directions of priority research in this area. They summarized previous studies involving threats such as impersonation, replay, speech synthesis, voice conversion, and spoofing attacks. The survey also presented recent efforts to develop dedicated countermeasures, together with some recommendations, such as the lack of standard datasets and the over-fitting of existing countermeasures to specific, known spoofing attacks.

2.7 Published works in the year 2014

Signal inconsistency due to environmental and acquisition channel factors is imposing a practical difficulty on speaker recognition. The noise affecting the voice signal varies greatly and a priori noise model is usually unavailable. To deal with this issue, Govindan *et al.* proposed a speaker recognition procedure that employs an adaptive wavelet shrinkage method for noise suppression, where wavelet sub-band coefficient thresholds, which are proportional to the noise contamination, are automatically computed [33].

A text-independent speaker recognition system was proposed by Madikeri. This work used the conventional i-vector modelling and a hybrid Probabilistic Principal Component Analysis (PPCA), which was tested by the total variability space (TVS) [34]. The research stated that the adopted approach showed a considerable decrease in development time, while the time required for training and testing remained unchanged. Speaker recognition performances were studied on the telephone conditions of the benchmark NIST SRE 2010 dataset with systems built on the Mel Frequency Cepstral Co-efficient (MFCC) feature.

Hernandez-Sierra *et al.* [35] proposed the use of a binary matrix to represent a speech extract. In this matrix, each acoustic frame was represented by a binary vector. The proposed representation relied on the Universal Background Model (UBM) paradigm but it shifts the speaker recognition workspace from a continuous probabilistic to a discrete binary space, allowing easy access to the speaker discriminant information. Additionally, the research proposed new variability compensation method in order to remove the unwanted attributes of session variability and the common attributes among speakers. Experimentations showed an Equal Error Rare (EER) improvement from 42% to 61%.

In a research presented by Smith, it was stated that much of the acoustic variation between the voices of men and woman is due to changes in the anatomical mechanisms for producing speech [36]. In his study, he proposed a measure to the duration required to discriminate whether a brief vowel segment was spoken by a man or woman, as well as the duration needed to correctly recognize what vowel was spoken [36]. The results from this research showed that reliable vowel recognition preceded reliable judgement on speaker sex.

Kreitewolf *et al.* [37] presented a work in which they tried to clarify the role of left and right hemispheres in the neural processing of linguistic prosody by using two functional magnetic resonance imaging (fMRI) experiments. In one experiment, the researchers controlled for stimulus influences by employing a prosody and speech task using the same speech material. In the second experiment, however, it was investigated whether a left-hemispheric involvement occurs when linguistic prosody is contrasted against other non-linguistic processes (i.e., speaker recognition). The results showed that the processing of linguistic prosody involves both hemispheres. They proposed that recognition of linguistic prosody is based on an inter-hemispheric mechanism. The mechanism is involving both a right-hemispheric sensitivity to pitch information and a left-hemispheric dominance in speech processing.

Kernel methods, including kernel-based speaker verification, are powerful techniques that were applied to pattern recognition problems. Chen [38] proposed kernels, which were referred to as the LR-based kernels and being derived by the Likelihood Ratio [LR) test, in attempts to integrate kernel methods with the LR-based speaker verification framework while an LR is embedded in the kernel function. The proposed kernels were claimed to have advantages over the existing methods and they outperformed conventional speaker verification approaches.

Larcher *et al.* [39] presented an evaluation protocol for the three main parts of the RSR2015 database, which was developed by the Institute for Info COMM Research (I2R) in Singapore. The research also presented the results of two speaker verification systems: The HiLAM system, based on a three-layer acoustic architecture, and an i-vector/PLDA system. As far as the research community is concerned, this research was considered as a reference performance evaluation scheme of RSR2015 database.

The importance of stereo-based stochastic feature compensation (SFC) methods for robust speaker verification (SV) in mismatched training and test environments was explored by Sarkar and Rao [40]. They proposed the application of Gaussian Mixture Model (GMM)-based SFC methods in an SV framework for background noise compensation. Features, that were extracted from a speaker's noisy and clean speech utterance (stereo data), were used to build front end GMMs. The results from this research reveal that the proposed Speaker Verification (SV) systems outperformed baseline SV systems in mismatched conditions across all noisy background environments.

An overview of the effective utilization of multiple utterances for speaker enrolment from a practical viewpoint was presented by Rajan *et al* [41]. Expressions for the evaluation of the likelihood ratio for the multi-enrolment case, together with details on how to compute the required matrix inversions and determinants, were provided. The results obtained from this research indicated that multi-condition training is more effective in estimating the Probabilistic Linear Discriminant Analysis (PLDA) hyperparameters than it is for likelihood computation. Further results supported a conclusion that i-vector averaging is a simple and effective way to process multiple enrolment utterances.

A simplified and supervised i-vector modelling approach was presented by Li *et al.* [42]. The research suggested applications to robust and efficient language identification and speaker verification. The supervised i-vectors were optimized such that they reconstruct the mean super vectors and minimize the mean square error between the original and the reconstructed label vectors. Factor analysis (FA) was performed on the pre-normalized centred GMM first order statistics super vector. A global table of the resulting matrices against the frame numbers' log values was constructed.

2.8 Published works in the year 2013

For the multi-point conference, a method for dominant speaker identification was proposed by Volfin and Cohen [43]. The motivation of this research is the need for reducing the amount of information that flows through the system during a multi-point conference, where routing and processing of the audio-visual information is very demanding on the network. The proposed approach assumes the use of speech activity information from time intervals of different lengths. The results reported about this system indicated reduction in the number of false speaker switches and improved robustness to transient audio interferences.

It is technically accepted that, at neutral talking environments, speaker recognition systems perform almost ideal, while a low performance of such systems is exhibited in emotional talking environments. This fact has motivated Shahin to present his research, which is based on investigating the issue by testing a database composed of 50 speakers talking in six different emotional states [44]. Experiments adopting speakers on neutral, angry, sad, happy, disgust, and fear sates showed improvement rates on speaker identification performance with 5.61%, 3.39%, and 3.06% compared to other published models.

Talkers, who may try to protect privacy and to avoid certain content from being overheard or made public, may use whispered speech as an alternative to the normal neutral speech. Automatic speaker identification systems trained with neutral speech may face a degraded performance in identifying a whispered speech. A feature transformation method was presented by Fan and Hansen [45]. The proposed method leads to identifying the speaker ID on whispered speech without using whispered adaptation data from test speakers. Three estimation methods were applied, including convolutional transformation (ConvTran), constrained maximum likelihood linear regression (CMLLR), and FA.

Pekhovsky and Sizov presented a comparison of speaker verification systems based on mixtures of probabilistic linear discriminant analysis (PLDA) models with Gaussian priors in a total variability space for speaker verification [46]. While considering the limitations of training database sizes, the research further analysed the conditions under which this application is advantageous. The results presented in this research indicated that a combination of a homogeneous i-vector extractor and a mixture of two Gaussian PLDA models is more effective than a cross-channel i-vector extractor with a single Gaussian PLDA.

Automatic speaker verification systems are facing the challenge of detecting the effects of vocal ageing. Kelly *et al.* presented their work in which they used a stacked classifier framework as a solution to speaker verification across ageing, by combining ageing and quality information with the scores of a baseline classifier [47]. An evaluation of a baseline Gaussian Mixture Model–Universal Background Model (GMM–UBM) system on a suitable database showed a progressive degradation in genuine speaker verification scores as ageing progresses.

In rapid detection and tracking strategy, a current speaker is required to be identified as an accepted member of an enrolled in-set group or rejected as an out-of-set speaker. Hansen *et al*. proposed a scoring algorithm that combined log likelihood scores across an energy-frequency grid [48]. In this model, the high-energy speaker dependent frames were fused with weighted scores from low-energy noise dependent frames. The research stated that keeping a balance between the speaker versus the background noise environment helped in realizing an improvement in the overall equal error rate (EER) performance.

Kua *et al.* [49] presented a paper in which an i-vector based sparse representation classification (SRC) was proposed as an alternative classifier to support vector machine (SVM) and Cosine Distance Scoring (CDS) classifier. The proposed classifier allowed the supports to be adapted to the test signal being characterized and, further, it did not require a training phase.

An automatic speaker age and gender identification approach was proposed by Li *et al*. [50]. The presented approach combined seven different methods at both acoustic and prosodic levels to improve the baseline performance. Additionally, four subsystems were proposed, which were demonstrated to be effective and able to achieve competitive results in classifying different age and gender groups. To further improve the overall classification performance, weighted summation-based fusion of these seven subsystems at the score level was exhibited.

The effects of training and test data duration and speaker's gender on the performance of speaker recognition systems was analysed by Hanilçi and Ertaş. Four conventional classifiers were used for speaker recognition [51]. The experiments were conducted on NIST 2002 and NIST 2005 speaker recognition evaluation (SRE) databases. The results indicated that recognition performance degraded when short utterances were used for training and testing data. Authors of this paper stated that recognition rate was found to be independent from the recognizer (e.g., equal error rate (EER) reduces from 10.33% to 27.86% on NIST 2005) and GSV–SVM system yields higher EER than other methods in the case of using short utterances.

Note that the lower the equal error rate (EER) value, the higher the accuracy of the system.

Ergonomic constraints and limited amount of computing resources were among the motivations of Larcher *et al*. while presenting their study about speaker recognition engines working on mobile devices [52]. Such systems may show efficient performance in classical context; however, their limitations will appear when restricting the quantity of speech data. To overcome this limitation, harnessing of the temporal structure of speech, using client-customized pass-phrases and new Markov model structures, was assumed as a suitable solution.

A combination of modified linear prediction coding (LPC) with wavelet transform (WT) for speaker identification, referred to as (AFLPC), was proposed by Daqrouq and Al Azzawi [53]. The distinguished speaker's vocal tract characteristics were extracted using the AFLPC technique and the size of a speaker's feature vector was optimized by means of genetic algorithm (GA). Because of its rapid response and ease in implementation, the Probabilistic Neural Network (PNN) was applied for classification. The results of this research stated that the PNN classifier achieves a better recognition rate (97.36%), by using the wavelet packet (WP) along with an AFLPC that is termed WPLPCF feature extraction method.

In the work presented by Sekar, the speech was converted into spectrogram, where an efficient representation of the speech signal in the form of pattern was utilized [54]. Image processing techniques were applied for the analysis and implementation of a text independent speaker identification system, where Radon Transform (RT) and Discrete Cosine Transform (DCT) were used to extract the features. The algorithm was tested and evaluated and the effect of number of Radon projections and DCT coefficients were analysed.

2.9 Published works in the year 2012

A new method of extracting the speech feature parameters for nonlinear and non-stationary signal based on the Hilbert-Huang transform (HHT) algorithm was presented by Liu *et al.* The speaker identification system was designed based on the Vector Quantization (VQ) [55]. Experiments on the system were carried out at different situations and showed that the system is feasible for speaker recognition.

Some speakers may possibly be linked with a field of expertise, like broadcast news or parliamentary speeches. Baum explored how topic information for a segment of speech, extracted from an automatic speech recognition transcript, can be employed to identify the speaker [56]. The researcher identified two methods for modelling topic preferences, one is based on speaker-characteristic keywords and the other considered automatically derived topic models to be assigned as topics to the speech segments. The proposed methods were tested on political speeches given in the German parliament by 235 politicians and found that topic signs do carry speaker information.

Some hardcore processors are having an embedded speech recognition system, which normally requires a considerably long time on train and recognition. Li *et al*. [57] addressed this issue and presented an FPGA-based speech recognition system implementations platform with the principle of vector quantization. The parallel hardware structure of the proposed system was implemented and tested. The results indicated a reduction in the time consumed for the training and recognition processes.

In order to match the noisy speech statistics to the clean speech, and to get a robust automatic speech and speaker recognition in noisy acoustic scenarios, feature coefficients are normalized. Squartini *et al.* stated that Histogram Equalization (HEQ) proved to be an effective normalization and transformation algorithm [58]. In their research, the presence of multi-channel acoustic channels was used to enhance the statistics modelling capabilities of the HEQ algorithm.

Krishnamoorthy *et al.* presented their work, in which they proposed that under limited data condition, the speaker recognition performance may be improved by controlled noise addition [59]. The problem of limited data for training and testing was resolved by noise values being added at very high Signal to Noise Ratio (SNR), where the added noise values caused an increase in the number of feature vectors being viewed as different instances of the given data. Research results indicated a performance of 78.20% with the use of limited data, and 80% using both limited and noisy data.

2.10 Published works in the year 2011

A comparison, in terms of performance between the Discrete Fourier Transform (DFT) and the Discrete Wavelet Transform (DWT), was presented by Turner *et al.* [60]. Both algorithms are widely applied with speaker recognition systems for extracting features from raw speech to capture the unique characteristics of a certain speaker. The results of this study emphasized a fact that DWT is favoured over the DFT in a wide variety of applications.

The T- and Z-norm score normalizations, that are widely used in speaker verification systems, require selection of a set of utterances in order to estimate the impostor score distribution. Apsingekar *et al.* [61] presented a study in which they investigated basing the utterances selection on speaker model clusters. They further proposed three normalization techniques, namely the Δ-, ΔT- and TC-norm. The results of their study indicated that it was possible to lower the equal error rate and minimum decision cost function.

Sadıç *et al.* [62] presented the common vector approach (CVA) to be used for text-independent speaker recognition. They further compared the performance of this approach with Fisher's Linear Discriminant Analysis (FLDA) and Gaussian Mixture Models (GMM). The results indicated that CVA has advantages in terms of processing power and memory requirement.

A speaker-independent hidden Markov model (HMM) - based voice conversion technique was proposed by Nose and Kobayashi. The study included context-dependent prosodic symbols obtained using adaptive quantization of the fundamental frequency (F0) [63]. The input utterance of a source speaker was decoded into phonetic and prosodic symbol sequences and the converted speech was generated using the decoded information. The promising results of this study for Japanese speech demonstrated that the adaptive quantization method gives better F0 conversion performance than the conventional one.

Automatic Speaker Verification (ASV) has received a lot of attention in recent years. Optimizing the dimensionality of feature space by selecting relevant features was the motivation of a study presented by Nemati and Basiri [64]. The proposed feature optimization method of this study was based on ant colony optimization (ACO) algorithm. In the method, and after feature reduction phase, feature vectors were applied to a Gaussian mixture model. The experimental results indicated that, with the optimized feature set, the performance of the ASV system was improved, and the speed of verification was significantly increased.

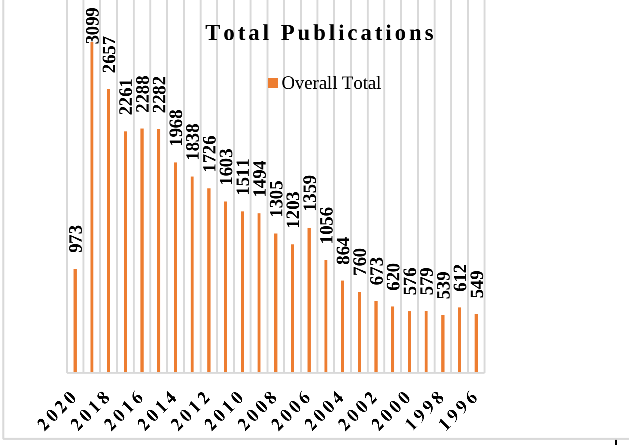
The impact of three special forms of the Minkowski metric on the performance of the conventional vector quantization (VQ) and Gaussian mixture model (GMM)-based closed-set text-independent speaker recognition systems was evaluated in a study presented by Hanilçi and Ertaş [65]. In addition to the evaluation performed in terms of recognition rate and confidence on decisions, the study made a comparison of results obtained from evaluations on clean speech and telephone speech databases.

3. Research Methodology

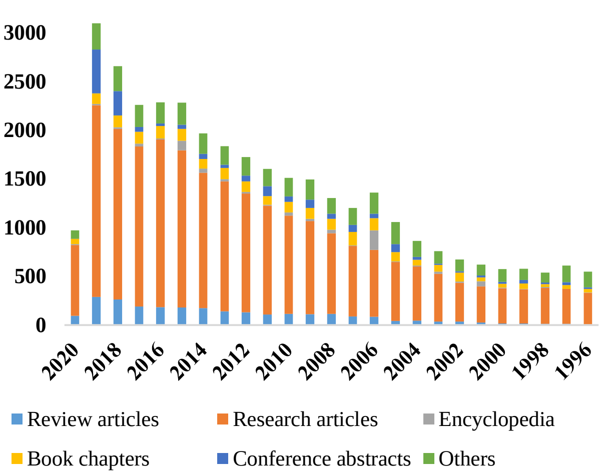
The research methodology, based on the adopted keywords, is divided up into three parts. Graphs illustrating a comparison, in terms of numbers, between published works in speaker identification, published works in speaker recognition, and published works in speaker verification, are presented in the first three subsections. While, the fourth subsection is considering the research articles being the majority in their numbers as per the graphs. An analysis in a tabular form is shown and, for each considered paper in the analysis, the year of publication, authors, the dataset and size of the dataset, the methodology adopted, as well as the accuracy of the obtained results, are presented in the tables.

3.1 Speaker Identification

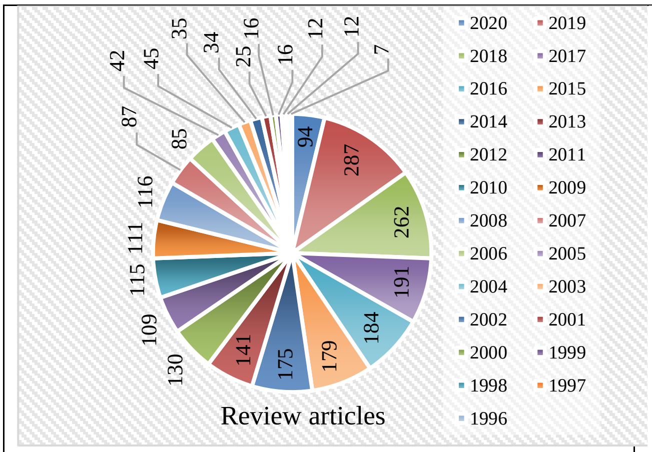
The search was conducted until March 2020 for the relevant works published within a period of 25 years between 1996 to 2020. The keywords "speaker identification" were used and the search revealed a total number of published works being 34395 items. This number of published items was found to be gradually increasing through the included years, from 1996 to 2020, as shown in Figure-2. Research articles are forming the majority among the searched items with a number of 22869 articles. One considerable observation is that the published research articles at the last two years, for the adopted keywords, exceeds 2000 items.



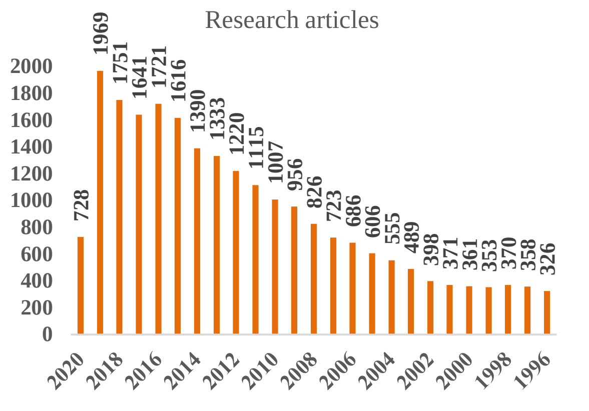
a) Total Publications



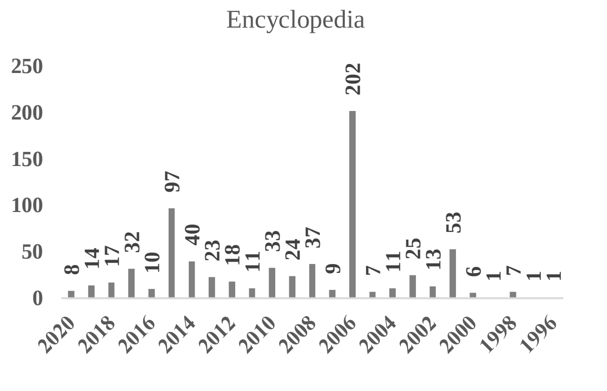
b) Publications Category



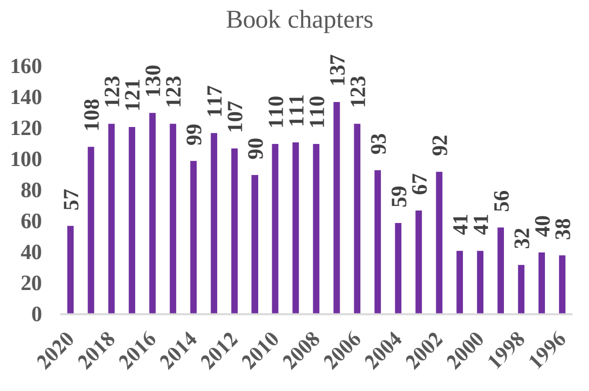
c) Review Articles



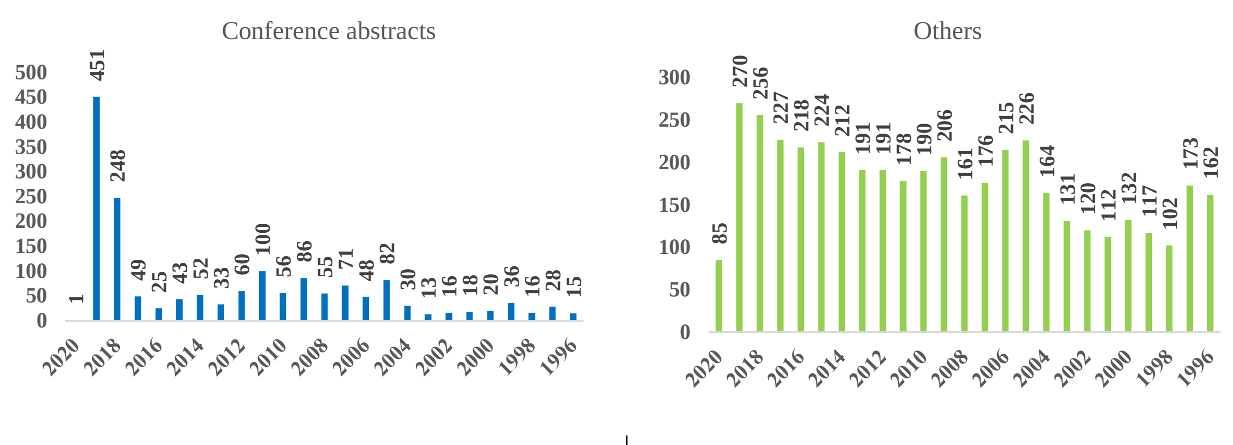
d) Research Articles



e) Encyclopedia



f) Book Chapters



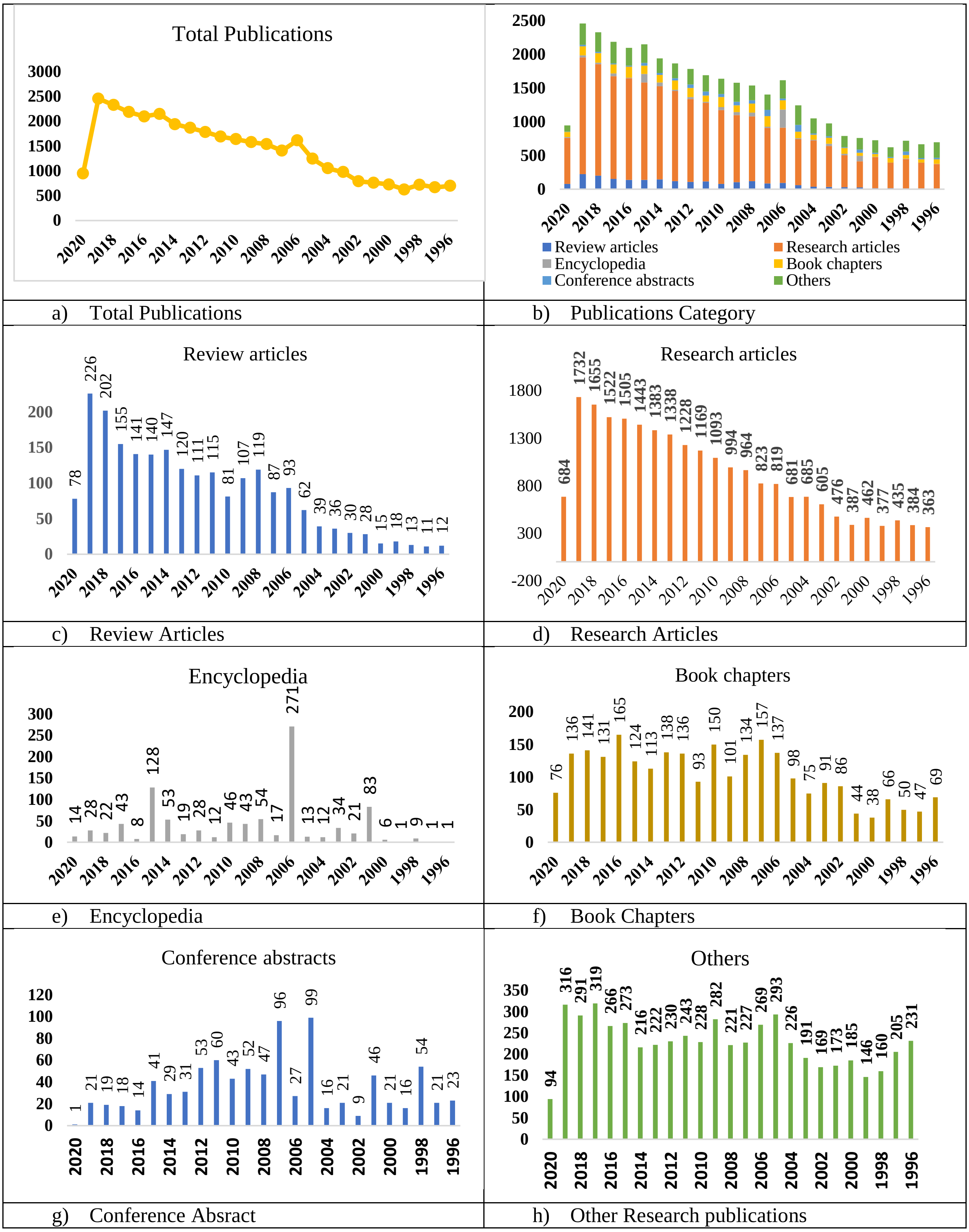
g) Conference Absract

h) Other Research publications

Figure -2 Published works in Science Direct Database about speaker identification for the years (1996-2020)

3.2 Speaker Recognition

During the same time (i.e. March 2020), we also performed a similar search for the same period of 25 years (1996 to 2020). The keywords used this time are “speaker recognition”. The search came up with a total of published works being 35510 items. In a similar trend, the number of publications is also increasing through the years from 1996 to 2020. This fact is shown in Figure-3, together with another observation about the number of research articles being the highest among other items. The total number of published research articles is 23207 and the number of articles published within the last two years alone is more than 2000 research articles.



a) Total Publications

b) Publications Category

c) Review Articles

d) Research Articles

e) Encyclopedia

f) Book Chapters

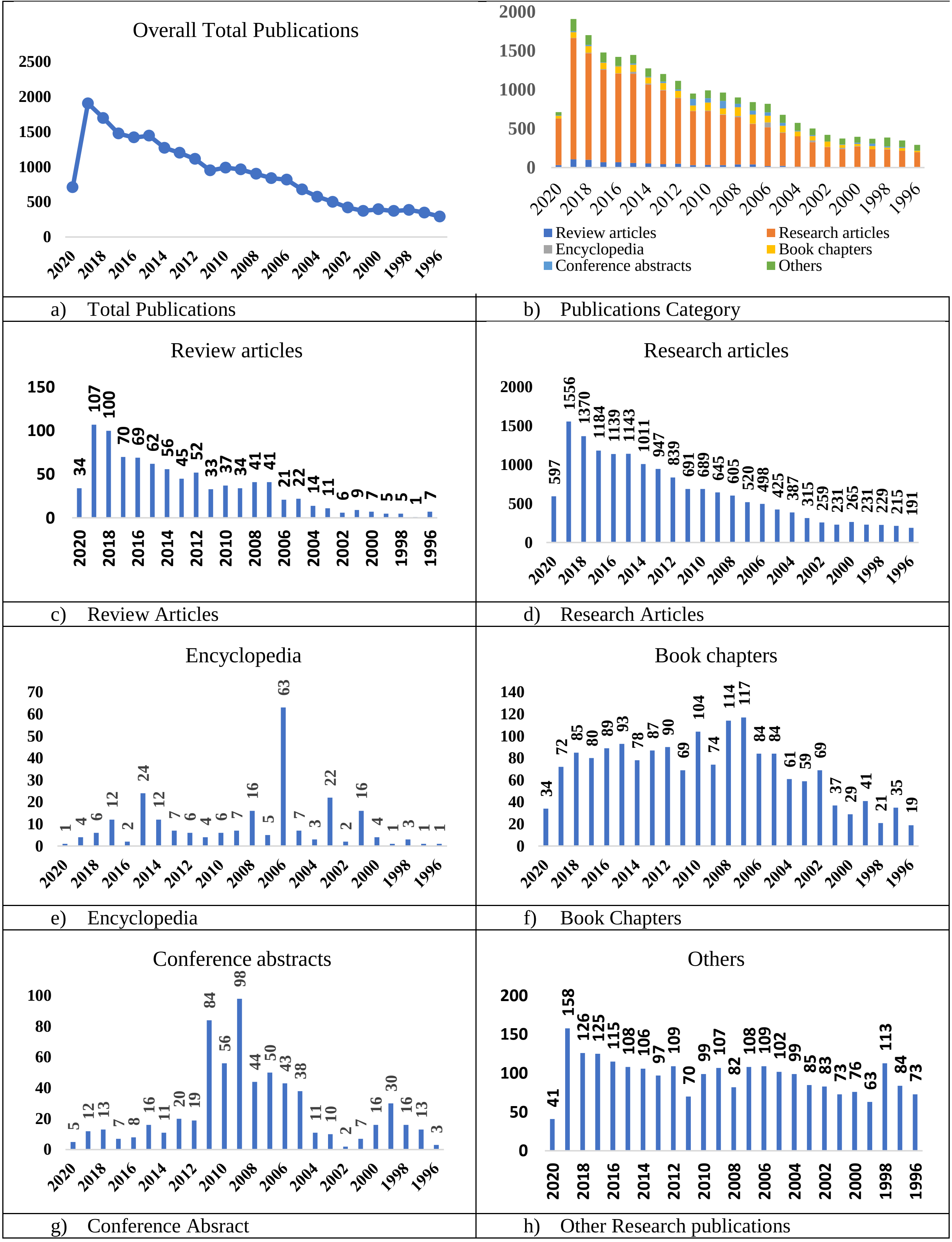
g) Conference Absract

h) Other Research publications

Figure -3 Published works in Science Direct Database about speaker recognition for the years (1996-2020)

3.3 Speaker Verification

As far as the search for published work in the field of speaker verification is concerned, a search was done on the month of March 2020 for the 25 years ranging between 1996 to 2020. The keywords used this time are "speaker verification" and the total published works that came out of the search is 22074 items. In harmony with the other two searches, the number of published items is in a gradual increase throughout the considered years (1996 to 2020), as shown in figure 4. Moreover, a noticeable part of this search is occupied by the research articles which are forming the majority, with a total number of 16182 articles. The published research articles at the last two years are about 1500. For the year 2020, however, the above data are collected for the first three months only.



a) Total Publications

b) Publications Category

c) Review Articles

d) Research Articles

e) Encyclopedia

f) Book Chapters

g) Conference Absract

h) Other Research publications

Figure -4 Published works in Science Direct Database about speaker verification for the years (1996-2020)

4. Comparing the Published Research Articles

The most important weight of the published works is focusing on research articles. Thus, in this section, we try to compare the published research articles within the key words "speaker identification, speaker recognition and speaker verification" in the period between 1996 to 2020. Figure 5 shows the comparison of the number of published research articles between these three items, in which it increases gradually between 1996 to 2020. In addition, both speaker identification and speaker recognition have the most effective weights of the published articles. It should be considered that, for 2020, the search was conducted only for the first half of the year. From this search, we can identify that the published articles on speaker recognition have the most effective rank, followed by speaker identification and then speaker verification.

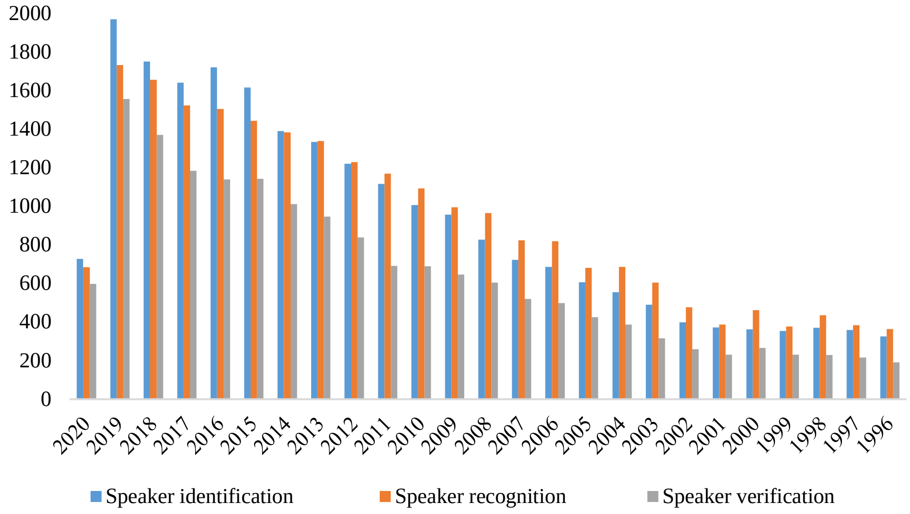


Figure -5 Comparing the published research articles in Science Direct Database for the years (1996-2020)

In the beginning of the time period covered by this survey, less research was conducted on speaker verification as compared to speaker identification and verification. In the period from 1996 till 2011, speaker recognition was an attractive area for researchers as compared to identification and verification. Later, the research focus and attraction was diverted to the area of identification, a trend which is still going on. It is worth mentioning that that above data includes the first three months only of the year 2020.

5. Analysis of the Published Research

As stated before, we divided the search into the three categories of speaker identification, speaker recognition, and speaker verification, which are demonstrated in tables 1, 2, and 3 respectively. Table 1 shows five published researches focusing on speaker identification, in which different methods were applied on different databases, attaining accuracy values between 81 and 97 %.

Table 1 - Extracted summary of five published researches relating to speaker identification

Ref. No.	**Authors**	**Year**	**Dataset**	**Data size (Speakers)**	**Method / Technique**	**Accuracy**
1	Ilana Volfin *et al.*	2013	TIMIT	-----	Binomial and Bin-Seq	-----
2	Ismail Shahin	2013	Collected	50	CSPHMM2s	81.50%
3	Xing Fan *et al.*	2013	TIMIT	20	ConvTran, CMLLR, FA	Up to 88.87%
4	Khaled Daqrouq et al.	2012	GMM	47	Modified LPC & WT	97.36%
5	K. Sekar	2012	BCS database	5	RT & DCT	96%

Table 2 shows thirty-two published researches focusing on speaker recognition, in which different methods were applied on different databases, reaching accuracy values between 80 and 100 %. TIMIT database was the most applied database and MFCC was the most used method.

Table 2- Extracted summary of thirty-two published researches relating to Speaker Recognition

Ref. No.	**Authors**	**Year**	**Dataset**	**Data size (Speakers)**	**Method / Technique**	**Accuracy**
6	Jesús Villalba *et al.*	2020	VoxCeleb 1,2	1000	Different methods	-----
7	Emma Jokinen *et al.*	2019	Private database	22	MFCC	80.3%
8	Ouassila Kenai *et al.*	2019	NOISEX database	400	MFCC	98%
9	Haojun Ai et al.	2019	Private database	10	hidden Markov model	98%
10	Stallone B. Sabatier *et al.*	2019	Collected twins database	167	GMM	-----
11	Ville Vestman *et al.*	2018	TIMIT data	-----	time-varying linear prediction	Improved 9%
12	Ankur Maurya *et al.*	2018	Private database	15	MFCC-GMM	94%
13	Noor Salwani Ibrahim *et al.*	2018	USM database	2656 bio-acoustics	Dimensionality Reduction	91%
14	Ing-Jr Ding et al.	2017	Collected database	10	SVD, GMM, DTW	85%
15	Mansour Alsulaiman *et al.*	2017	-----	-----	survey	Up to 94%
16	Suma Paulose *et al.*	2017	TIMIT database	630	MFCC	89%
17	Eleonora D'Arca *et al.*	2016	Private database	-----	MFCCs	94%
18	Anzar S.M. *et al.*	2016	ELSDSR & ELDASR	50	VQ & GMM	93%
19	Sharada V. *et al.*	2015	MVSR & Hindi	100	Normalized Dynamic Spectral Features	Up to 100%
20	Fanhu Bie *et al.*	2015	NIST SRE2008	51	GMM–UBM	80%
21	Sumithra Manimegalai Govindan *et al.*	2014	TIMIT King database	630	ABWS	80%
22	Srikanth R. Madikeri	2014	NIST SRE 2010	-----	hybrid FA/PPCA	70%

23	Gabriel Hernandez-Sierra *et al.*	2014	NIST SRE 2005	124	LDA-WCCN	improved 61%
24	David R. R. Smith	2014	Collected data	50	speaker-sex discrimination	75%
25	Jens Kreitewolf *et al.*	2014	MRI data	19	LCD projector	92.67%
26	John H. L. Hansen *et al.*	2013	TIMIT	60	frequency partitioning	EER improved 10.8%
27	Ming Li *et al.*	2013	Different	-----	MFCC, SVM, GMM, SVM + 450-dimensional	improved up to 5.6%
28	Cemal Hanilçi *et al.*	2013	NIST 2005	616	GMM–UBM, VQ–UBM, SVM–GLDS GSV–SVM	reduced up to 27.86%
29	Liwei Liu *et al.*	2012	Collected data	40	HHT-IF and LPCC	Up to 100%
30	Doris Baum	2012	Collected data	253	SVM and LDA	Improved EER 8.6%
31	Jingjiao Li *et al.*	2012	-----	-----	MFCC and VQ	93.3%
32	Stefano Squartini *et al.*	2012	FAK_5Aoft heAurora2	104	Histogram equalization	81.08%
33	P. Krishnamoorthy *et al.*	2011	TIMIT	100	MFCC and GMM	Up to 80%
34	Claude Turner *et al.*	2011	TIMIT	16	DWT & DFT	improved
35	Selami Sadıç *et al.*	2011	TIMIT	20	CVA and GMM	Up to 100%
36	Takashi Nose *et al.*	2011	ATR Japanese speech	-----	HMM	Up to 92%
37	Cemal Hanilçi *et al.*	2011	TIMIT NTIMIT	630 168	VQ/GMM	Up to 70%

Table 3 shows twenty-nine published researches focusing on speaker verification, in which different methods were applied on different databases, reaching EER improvement between 3 and 15 %. NIST SRE database was the most applied database and both GMM and UBM were the most used methods.

Table -3 Extracted summary of twenty-nine published researches relating to speaker verification

Speaker Verification

Ref. No.	**Authors**	**Year**	**Dataset**	**Data size (Speakers)**	**Method / Technique**	**Accuracy**
38	Ville Vestman *et al.*	2020	VoxCeleb2	7365	Different methods	-----
39	Soufiane Hourri *et al.*	2019	THUYG-20 SRE	-----	MFCC	92%
40	Amos Treiber *et al.*	2019	-----	3000 subjects	PLDA	high
41	Kuruvachan K. George *et al.*	2018	NIST 2004 SRE	2600	CDF-SVM	improved 4.5%
42	M. S. Athulya *et al.*	2018	TIMIT database	630	modified PNCC	Error reduced 15%
43	Rosa González *et al.*	2017	self-collected	60	trial-by-trial decisions	81%
44	Milton Sarria-Paja *et al.*	2017	CHAIN IMIT	462 476	MFCCs	66%
45	Rania M. Ghoniem *et al.*	2017	Private database	500 signals	FHMM	98%

46	Linlin Wang *et al.*	2016	created speech database	-----	MFCC	high
47	Ming Li *et al.*	2016	X-ray Microbeam database RSR 2015 database	46	MFCCs	Error reduction 15%
48	Jesús Villalba *et al.*	2016	NIST SRE10	-----	MFCC	Error reduction 14%
49	Wei Rao *et al.*	2016	NIST 2012 SRE	1931	PLDA–RVM	high
50	Yuan Liu et al.	2015	RSR2015	194	GMM-UBM	Error 0.1
51	Rosa González Hautamäki *et al.*	2015	Collected data	34 trials	GMM-UBM	Error 10%
52	B. C. Haris *et al.*	2015	NIST 2012 SRE	1931	GMM-SV	Improved 7%
53	Yong Wang *et al.*	2015	-----	-----	GMM–UBM	Improved 3%–4%.
54	Zhizheng Wu *et al.*	2015	-----	-----	survey	-----
55	Yi-Hsiang Chao	2014	XM2VTSDB	30	LR-based discriminant kernel	-----
56	Anthony Larcher *et al.*	2014	RSR2015	194	GMM & HMM	high
57	Sourjya Sarkar *et al.*	2014	NIST-2003-SRE	356	GMM-based SFC	EER improved 3.07%
58	Padmanabhan Rajan *et al.*	2014	NIST 2012 SRE	1931	PLDA	Degradation of 8% EER
59	Ming Li *et al.*	2014	NIST SRE 2010 NIST LRE 2007	-----	GFCC	-----
60	Timur Pekhovsky et al.	2013	NIST's SRE 1998–2008	4329	cepstral mean subtraction	13% EER reduction
61	Finnian Kelly *et al.*	2013	Trinity College Dublin	18	GMM–UBM	
62	Jia Min Karen Kua *et al.*	2013	NIST 2010 SRE	500	i-vector based SRC	EER reduction 8–19%
63	Anthony Larcher *et al.*	2013	MyIdea database	900	GMM/UBM	Gain up to 65%
64	Vijendra Raj Apsingekar *et al.*	2011	NTIMIT and NIST-2002 corpora and compare	629	SMCs	improved EER 11.02%
65	Shahla Nemati *et al.*	2011	TIMIT corpora	630	ant colony optimization	improved EER

6. Discussion

Because speech is known to contain a lot of information about the speaker, it is found to be the best carrier of information in human communication systems. When people are talking to each other, they almost always find little difficulty to recognize speakers and even their emotional state. It is, however, a complex process when the process is implemented in a machine. Speech signal processing is the field of study being adopted to find methods that help machines to recognize speakers. Speaker recognition is broadly classified into speaker identification and speaker verification.

The three areas (i.e., speaker identification, speaker verification, and speaker recognition) are the main keywords around which our study is built.

A study that covers a quarter-century research outcomes is presented in this paper, where analysis of methods and techniques used in the three mutually related fields, which are speaker identification, speaker recognition, and speaker verification, is conducted. This exploration study was carried out by considering the research databases of the last 25 years from 1996 till 2020 using Science Direct databases. Due to the high number of published works in these fields, we can conclude that these three fields occupied an important part in the area of research. In addition, this work is focusing on research papers that give a high impact indication in this field. This work gives guidelines for researchers to select the best method applied on the best database to achieve high accuracy. Data relating to the year of 2020 is showing the first three months only.

Our rResults obtained in this review paper found that a tremendous research work was conducted in these areas and various techniques were adopted within various time and frequency domains. Features and improvements in the accuracy level and achievements in the experiments using the various techniques were pretested in a tabulated form, showing that selection of techniques was depending on the type of problem.

The problem of speaker recognition gained high focus from the beginning of the covered study period, compared with speaker verification and identification studies. The study of speaker verification typically focuses on improvements made on outcomes of previous studies, which gives this field a wider scope in the future with respect to the other two fields. This conclusion may encourage young researchers, who wish to work in the area of acoustic or speech technology, to consider areas as speaker verification with higher priority. The results of the collected and analysed data reflected the significant of speech processing research throughout the world. It was found that all the three considered fields are highly active research areas with an average of at least 10-15 research activities per day.

A quick view to the three graphs produced during the analysis presented in this paper indicates that the sum of published research articles is outnumbering other categories of publications (i.e. review articles, encyclopaedia, book chapters, conference abstracts, and others). Furthermore, the graphs introduced additional important information, including the dataset used, the size of the data adopted, the implemented methods, and the accuracy of the obtained results in the analysed research, which are extracted from the explored publications.

7. Conclusions

Three interrelated scientific fields were considered in the critical review that is presented in this paper. Speaker identification, speaker recognition, and speaker verification are the research areas that are adopted, where a systematic analysis for the methods and techniques used in the establishment of their theoretical basis and applicability was conducted. Research outcomes for a span of quarter-century was investigated and presented by searching through the well-recognized Science Direct database.

During the years 1996 until 2020, huge research work was conducted and numerous techniques were adopted. For each of the three areas, the examined published works were categorized as review articles, research articles, encyclopaedia, book chapters, conference abstracts, and others. Summaries of these kinds of literature and plots showing the number of examined work were included in this paper. Additionally, statistical analyses representing the publications and their categories over the mentioned period were illustrated in tables with important extracted information, such as the dataset used, the size of the data adopted, the implemented methods, and the accuracy of the obtained results in the analysed research. One important point to conclude is that the number of researches in voice identification, recognition, and verification shows increased trend, with most of these publications reached to a significantly high level of accuracy in their findings.

Conflicts of Interest: None.

- We hereby confirm that all the Figures and Tables in the manuscript are ours. Besides, the Figures and images which are not ours have been given the permission for re-publication, attached with the manuscript.
- Ethical Clearance: No issue in this regard.

References

1. V. Vestman, T. Kinnunen, R. Hautamäki, and G.M. Sahidullah, GM. **2020**. "Voice Mimicry Attacks Assisted by Automatic Speaker Verification," *Computer Speech and Language*, **59**(1): 36–54.

2. J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, and N. Dehak. **2020**. "State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and Speakers in the Wild evaluations," *Computer Speech and Language*, **60**(4): 101026.

3. E. Jokinen, R. Saeidi, T. Kinnunen, and P. Alku. **2019**. "Vocal effort compensation for MFCC feature extraction in a shouted versus normal speaker recognition task," *Computer Speech and Language*, **53**(1): 1–11.

4. O. Kenai, S. Djeghiour, N. Asabai, and M. Guerti. **2019**. "Forensic Gender Speaker Recognition under Clean and Noisy Environments," *Procedia Computer Science*, **151**: 897-902.

5. S. Hourri, and J. Kharroubi. **2019**. "A Novel Scoring Method Based on Distance Calculation for Similarity Measurement in Text-Independent Speaker Verification." *Procedia Computer Science*, **148**: 256–265.

6. H. Ai, K. Tang, L Han, Y. Wang, and S. Zhang. **2019**. "DuG: Dual speaker-based acoustic gesture recognition for humanoid robot control," *Information Sciences*, **504**(12): 84–94.

7. S. Sabatier, M. Trester, and J. Dawson. **2019**. "Measurement of the impact of identical twin voices on automatic speaker recognition," *Measurement*, **134**(2): 385-389.

8. A. Treiber, A. Nautsch, J. Kolberg, and Schneider. **2019**. "Busch C. Privacy-preserving PLDA speaker verification using outsourced secure computation," *Speech Communication*, **114**(11): 60–71.

9. K. George, C. Kumar, S. Sivadas, K. Ramachandran, and A. Panda. **2018**. "Analysis of cosine distance features for speaker verification," *Pattern Recognition Letters*, 112(9): 285–289.

10. V. Vestman, D. Gowda, M. Sahidullah, P. Alku, and T. Kinnunen. **2018**. "Speaker recognition from whispered speech: A tutorial survey and an application of time-varying linear prediction," *Speech Communication*, **99**(5): 62–79.

11. A. Maurya, D. Kumar, and R. Agarwal R. **2018**. "Speaker Recognition for Hindi Speech Signal using MFCC-GMM Approach," *Procedia Computer Science*, **125**: 880–887.

12. M. Athulya, and P. Sathidevi. **2018**. "Speaker verification from codec distorted speech for forensic investigation through serial combination of classifiers," *Digital Investigation*, **25**(6): 70–77.

13. N. Ibrahim, and D. Ramli. **2017**. "I-vector Extraction for Speaker Recognition Based on Dimensionality Reduction," *Procedia Computer Science*, **126**: 1534–1540.

14. J. Ding, and J. Shi. **2017**. "Kinect microphone array-based speech and speaker recognition for the exhibition control of humanoid robots," *Computers & Electrical Engineering*, **62**(8): 719–729.

15. R. Hautamäki, M. Sahidullah, V. Hautamäki, and T. Kinnunen. **2017**. "Acoustical and perceptual study of voice disguise by age modification in speaker verification," *Speech Communication*, **95**(12): 1–15.

16. M. Alsulaiman, A. Mahmood, G. Muhammad. **2017**. "Speaker recognition based on Arabic phonemes," *Speech Communication*, 86(2): 42–45.

17. M. Sarria-Paja, and T. Falk. **2017**. "Fusion of auditory inspired amplitude modulation spectrum and cepstral features for whispered and normal speech speaker verification" *Computer Speech & Language,* **45**(9): 437–456.

18. R. Ghoniem, and K. Shaalan. **2017**. "A Novel Arabic Text-independent Speaker Verification System based on Fuzzy Hidden Markov Model," *Procedia Computer Science*, **117**: 274–286.

19. S. Paulose, D. Mathew, and A. Thomas. **2017**. "Performance Evaluation of Different Modeling Methods and Classifiers with MFCC and IHC Features for Speaker Recognition" *Procedia Computer Science*, **115**: 55–62.

20. L. Wang, J. Wang, L. Li, T. Zheng, and F. Soong. **2016**. " Improving speaker verification performance against long-term speaker variability," *Speech Communication*, **79**(5): 14–29.

21. E. D'Arca, N. Robertson, and J. Hopgood. **2016**."Robust indoor speaker recognition in a network of audio and video sensors," *Signal Processing*, **129**(12): 137–149.

22. M. Li, J. Kim, A. Lammert, P. Ghosh, and S. Narayanan. **2016**. "Speaker verification based on the fusion of speech acoustics and inverted articulatory signals," *Computer Speech and Language,* **36**(3): 196–211.

23. J. Villalba , A. Ortega, A. Miguel, and E. Lleida. **2016**."Analysis of speech quality measures for the task of estimating the reliability of speaker verification decisions," *Speech Communication*, **78**(4): 42–61.

24. W. Rao, and M. Mak. **2016**. "Sparse kernel machines with empirical kernel maps for PLDA speaker verification," *Computer Speech and Language,* **38**(7): 104–121.

25. S. Anzar, K. Amala, R. Rajendran, A. Mohan, and F. Aziz. **2016**. "Efficient online and offline template update mechanisms for speaker recognition," *Computers and Electrical Engineering*, **50**(2): 10–25.

26. S. Chougule, and M. Chavan. **2016**. "Robust Spectral Features for Automatic Speaker Recognition in Mismatch Condition," *Procedia Computer Science*, **58**: 272–279.

27. Y. Liu, Y. Qian, N. Chen, T. Fu, and K. Yu. **2015**. "Deep feature for text-dependent speaker verification," *Speech Communication*, **73**(10): 1–13.

28. R. Hautamäki, T. Kinnunen, V. Hautamäki, and A. Laukkanen. **2015**. "Automatic versus human speaker verification: The case of voice mimicry," *Speech Communication*, **72**(9): 13–31.

29. B. Haris, and R. Sinha. **2015**. "Low-complexity speaker verification with decimated supervector representations," *Speech Communication*, **68**(4): 11–22.

30. F. Bie, D. Wang, J. Wang, and T. Zheng. **2015**. "Detection and reconstruction of clipped speech for speaker recognition," *Speech Communication*, **72**(9): 218-231.

31. Y. Wang, H. Wu, J. Huang. **2015**. "Verification of hidden speaker behind transformation disguised voices," Digital Signal Processing, **45**(10): 84–95.

32. Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, and H. Li. **2015**. "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, **66**(2): 130–153.

33. S. Govindan, P. Duraisamy, and X. Yuan. **2014**. "Adaptive wavelet shrinkage for noise robust speaker recognition," *Digital Signal Processing*, **33**(10): 180–190.

34. S. Madikeri. **2014**. "A fast and scalable hybrid FA/PPCA-based framework for speaker recognition," *Digital Signal Processing*, **32**(9): 137–145.

35. G. Hernandez-Sierra, J. Calvo, J. Bonastre, and P. Bousquet. **2014**. "Session compensation using binary speech representation for speaker recognition," *Pattern Recognition Letters*, **49**(11): 17–23.

36. D. Smith. **2014**. "Does knowing speaker sex facilitate vowel recognition at short durations?," *Acta Psychologica*, **148**(5): 81–90.

37. J. Kreitewolf, A. Friederici, and K. Kriegstein. **2014**. "Hemispheric lateralization of linguistic prosody recognition in comparison to speech and speaker recognition," *NeuroImage*, **102**(11): 332–344.

38. Y. Chao. **2014**. "Using LR-based discriminant kernel methods with applications to speaker verification," *Speech Communication,* **57**(2): 76–86.

39. A. Larcher, K. Lee, B. Ma, and H. Li. **2014**. "Text-dependent speaker verification: Classifiers, databases and RSR2015", *Speech Communication*, **60**(5): 56–77.

40. S. Sarkar, and K. Rao. **2014**. "Stochastic feature compensation methods for speaker verification in noisy environments," *Applied Soft Computing*, **19**(6): 198–214.

41. P. Rajan, A. Afanasyev, V. Hautamäki, and T. Kinnunen. **2014**. "From single to multiple enrolment i-vectors: Practical PLDA scoring variants for speaker verification," *Digital Signal Processing,* **3**(8): 93-101.

42. Li M, Narayanan S. **2014**. Simplified supervised i-vector modeling with application to robust and efficient language identification and speaker verification. *Computer Speech & Language,* **28**(4): 940–958.

43. I. Volfin, and I. Cohen. **2013**. "Dominant speaker identification for multipoint videoconferencing," Computer Speech and Language, **27**(4): 895–910.

44. I. Shahin I. **2013**. "Speaker identification in emotional talking environments based on CSPHMM2s," *Engineering Applications of Artificial Intelligence,* **26**(7): 1652–1659.

45. X. Fan, and J. Hansen. **2013**. "Acoustic analysis and feature transformation from neutral to whisper for speaker identification within whispered speech audio streams," *Speech Communication,* **55**(1): 119–134.

46. T. Pekhovsky, and A. Sizov. **2013**. "Comparison between supervised and unsupervised learning of probabilistic linear discriminant analysis mixture models for speaker verification," Pattern *Recognition Letters,* **34**(1): 1307–1313.

47. F. Kelly, A. Drygajlo, and N. Harte. **2013**. "Speaker verification in score-ageing-quality classification space," *Computer Speech and Language,* **2**(5): 1068–1084.

48. J. Hansen, J. Suh, and M. Leonard. **2013**. "In-set/out-of-set speaker recognition in sustained acoustic scenarios using sparse data," *Speech Communication,* **55**(6): 769–781.

49. J. Kua, J. Epps, and E. Ambikairajah. **2013**. "i-Vector with sparse representation classification for speaker verification," Speech Communication, **55**(5): 707–720.

50. M. Li, K. Han, and S. Narayanan. **2013**. "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech and Language,* **27**(1): 151–167.

51. C. Hanilçi, and F. Ertaş. **2013**. "Investigation of the effect of data duration and speaker gender on text-independent speaker recognition," *Computers & Electrical Engineering,* **39**(2): 441–452.

52. A. Larcher, J. Bonastre, and J. Mason. **2013**. "Constrained temporal structure for text-dependent speaker verification," *Digital Signal Processing,* **23**(6): 1910–1917.

53. K. Daqrouq, and K. Al Azzawi. **2013**. "Average framing linear prediction coding with wavelet transform for text-independent speaker identification system," *Computers and Electrical Engineering,* **38**(6): 1467–1479.

54. K. Sekar. **2013**. "Performance Analysis of Text-Independent Speaker Identification System," *Procedia Engineering,* **38**: 1925–1934.

55. L. Liu, F. Qian, and Y. Zhang. **2012**. "Application Research of HHT-IF Speech Feature Parameter in Speaker Recognition System," *Energy Procedia,* **17**: 1102–1108.

56. D. Baum. **2012**. "Recognising speakers from the topics they talk about. *Speech Communication*," **54**(12): 1132–1142.

57. J. Li J, D. An, L. Lang, and D. Yang. **2012**. "Embedded Speaker Recognition System Design and Implementation Based on FPGA," *Procedia Engineering,* **29**: 2633–2637.

58. S. Squartini, E. Principi, R. Rotili, and F. Piazza. **2012**. "Environmental robust speech and speaker recognition through multi-channel histogram equalization," *Neurocomputing,* 78(1): 111–120.

59. P. Krishnamoorthy, H. Jayanna, and S. Prasanna. **2012**. "Speaker recognition under limited data condition by noise addition," *Expert Systems with Applications,* **38**(10): 13487–13490.

60. C. Turner, A. Joseph, M. Aksu, and H. Langdond. **2011**. "The Wavelet and Fourier Transforms in Feature Extraction for Text-Dependent, Filterbank-Based Speaker Recognition," *Procedia Computer Science,* **6**: 124-129.

61. V. Apsingekar, and P. De Leon. **2011**. "Speaker verification score normalization using speaker model clusters," *Speech Communication*. **53**(1): 110–118.

62. S. Sadıç, and M. Gülmezoğlu. **2011**. "Common vector approach and its combination with GMM for text-independent speaker recognition," *Expert Systems with Applications,* **38**(9): 11394-11400.

63. T. Nose, and T. Kobayashi. **2011**. "Speaker-independent HMM-based voice conversion using adaptive quantization of the fundamental frequency," *Speech Communication,* **53**(7): 973–985.

64. S. Nemati, and M. Basiri. **2011**. “Text-independent speaker verification using ant colony optimization-based selected features,” *Expert Systems with Applications*, **38**(1): 620–630.

65. C. Hanilçi, and F. Ertaş. **2011**. “Comparison of the impact of some Minkowski metrics on VQ/GMM based speaker recognition,” *Computers & Electrical Engineering*, **37**(1): 41–56.