Artificial Intelligence Review (2024) 57:58
https://doi.org/10.1007/s10462-023-10688-w

# Milestones in speaker recognition

R. Sharma[1] · D. Govind[2] · J. Mishra[3] · A. K. Dubey[2] · K. T. Deepak[1] · S. R. M. Prasanna[3]

Accepted: 30 December 2023 / Published online: 15 February 2024

## Abstract

This article reviews significant research in the domain of speaker recognition, i.e., the task of determining the speaker's identity from its speech. Unlike conventional review articles, this document strives to be concise and selective, provide a historical context, and reach a wider audience. In this endeavour, a summary of selected key works of every decade is provided which highlights the theme(s) of research of that period. At first, an overview of the humble beginnings of the 1960s and 70s is provided, followed by the key developments in the 80s and 90s. The prime focus of the research community in the 2000s is then discussed, leading to various non-conventional features, modelling techniques, and hybrid or fusion systems. The developments of the last decade (the 2010s), such as the i-vector-based systems, are then discussed. Modern speaker recognition based on Artificial Intelligence (AI), such as the x-vector system, and refinements of the i-vector-based systems using deep neural networks, are then discussed. The article concludes with a concise discussion of the evolving recent trends and allied research in speaker recognition.

**Keywords** Speaker recognition · Speaker identification · Speaker verification · Text-dependent · Text-independent · History · Review

✉ R. Sharma
rajibd2k@yahoo.com

D. Govind
d_govind@kluniversity.in

J. Mishra
jagabandhu.mishra.18@iitdh.ac.in

A. K. Dubey
dubey18oct@kluniversity.in

K. T. Deepak
deepak@iiitdwd.ac.in

S. R. M. Prasanna
prasanna@iitdh.ac.in

[1] Indian Institute of Information Technology (IIIT) Dharwad, Karnataka, India

[2] Koneru Lakshmaiah Education Foundation (KLEF), Vaddeswaram, Andhra Pradesh, India

[3] Indian Institute of Technology (IIT) Dharwad, Karnataka, India

## 1 Introduction

Speech is, and will always be, the principal mode of communication in human society. A speech signal contains a remarkable amount of information, apart from the obvious message that the speaker tries to convey, such as the gender, mood, health, and most importantly, the identity of the speaker. Just like one's DNA, one's voice is also unique. Add to that the fact that speech is possibly the easiest signal to record and transmit/receive: a mobile-phone connection is all that is required. It is the cheapest and the most widely accessible technology available in the world. This, inevitably, attracts a lot of attention to speech-based applications, one of which is to utilize speech as a biometric marker or 'voice-print'. This application has been traditionally called speaker recognition (Benesty et al. 2008).

The domain of speaker recognition is generally split into two sub-domains: speaker identification and speaker verification. Speaker identification is concerned with determining the speaker of a speech utterance from a set of speakers. It is a multi-class problem. Speaker verification, on the other hand, verifies the claim of a particular speech utterance belonging to a particular speaker. It is a two-class problem. Additionally, in a given speaker recognition system, the set of speakers may be closed (fixed and limited) or open (unlimited). Traditionally, each of the two sub-domains is further bifurcated into two categories: text-dependent and text-independent. In the text-dependent scenario, the speakers are allowed to utter only a pre-defined fixed set of phrases. In the text-independent scenario, there are no such restrictions: the speaker may speak whatever it wishes. Of course, most of the research work to date focuses on the English language, and more complicated scenarios involving other languages, dialects or accents are emerging recently. Nevertheless, the exact type of speaker recognition that is pursued depends on the end goal of the application, speaker verification (both text-dependent and text-independent) being the modus operandi in most cases (Furui 2005; Reynolds 2002; Faundez-Zanuy and Monte-Moreno 2005; Bai and Zhang 2021; Hanifa et al. 2021; Kabir et al. 2021).

The genesis of speaker recognition dates back to the early 60s, an era before the internet or the personal computer. From the slow progress of its early days, it has today become one of the most significant areas of research, which is likely to affect many areas of public and private life in the not-too-distant future. In this review work, we aim to tell the story of this field in a manner which can reach a wider audience, which is not overwhelming, and which can help budding researchers venture into the field more easily. At the same time, we believe that this article will be an excellent read for advanced researchers and experts, not just to take a walk down memory lane, but also to gather a summary of the recent trends in the field. In our quest for producing a quality review, we have hand-picked and summarized only a few research articles that showcase the mood of research in the decades leading up to the present. The number of citations, and the popularity of the papers in the research community have been the guidelines in the selection of these works. In the process, we have most certainly missed out on many important works of the past: but that is a trade-off we have had to make to suit our objective. To present the historical perspective, we found inspiration in the style of Furui's review paper published in 2004-05, and have tried to incorporate the same in our document (Furui 2005).

The rest of the document is organised as follows: Sect. 2 summarizes the research activities during the 1960s and 70s. Sect. 3 discusses the significant works conducted in the 1980s. Sects. 4, 5, 6 and Sects. 7, 8, 9 discuss the research highlights of the 1990s and 2000s, respectively. It is a testimony to the importance of the 1990s and the 2000s in the advancement of speaker recognition technology. Sect. 10 discusses the modern state of speaker recognition technology, from the 2010s - the present. Multiple sub-sections are included in this section devoted to diversified research activities in the present era. Sects. 11 and 12 summarizes and concludes this article, respectively.

## 2 Onset of speaker recognition: 1960s and 70s

The dawn of speaker recognition technology happened sometime in the early 1960s, a decade or so later than the advent of speech recognition. One of the first notable works was conducted by Pruzansky et. al., in the Bell Labs (Pruzansky and Mathews 1964). The experiments were conducted on a group of ten talkers uttering ten different words seven times each. The highlights of the work were:

- The system utilized a simple one-to-one matching of the features of the test-utterance with those of the model or train utterances of the closed-set of speakers.
- Different features were extracted by averaging the spectrogram in time and frequency using rectangular regions of varying sizes.
- The features which provided the best ‘F-ratio’, i.e., the ratio of inter-speaker variance to intra-speaker variance, was chosen as the best features for the speaker recognition task.
- Varying the rectangular area size was found to affect recognition performance. For a given number of features, successively more time averaging increased performance while averaging over frequency channels lowered recognition. This indicated that with respect to information about talker identity the energies in successive frequency channels are relatively independent but that the energies in successive time intervals are relatively more dependent.

Over time, many improvements were proposed to the Bell Labs system by many researchers. One notable contribution was by Doddington, who conducted a speaker-verification experiment using eight known speakers and 32 impostors (Doddington 1971). This was also one of the first works which defined and contrasted the speaker-verification problem with the speaker-identification problem. Formant frequencies, voicing pitch period, and speech energy were used as features and nonlinear time normalization was performed by maximizing the correlation between the sample and reference second-formant profiles through a piecewise linear continuous transformation of time.

Almost a decade after the birth of speaker recognition, researchers started seeking answers to fundamental questions such as the problem of intra-speaker variability and the utility and reliability of the spectrogram in speaker recognition (Bricker et al. 1971; Endres et al. 1971). Endres et al., for example, studied spectrograms of utterances produced by seven speakers and recorded over periods of up to 29 years (Endres et al. 1971). Following were the key observations of their work:

- The formants and pitch of voiced sounds shift to lower frequencies with increasing age of test persons.
- Spectrograms of texts spoken in a normal and a disguised voice revealed strong variations in formant structure.
- Spectrograms of utterances of well-known people were compared with those of imitators. The imitators were able to vary the formant structure and fundamental frequency

> of their voices but failed to adapt these parameters to match or even be similar to those of the imitated persons.

Up until this point, most of the research on speaker recognition was text-dependent, i.e., the same phrases were used during training and testing. One of the first works in text-independent speaker recognition was conducted by Atal, in the Bell Labs (Atal 1972). A speaker-identification experiment was performed on a population of 10 female speakers, and the segment of the utterance used to identify an individual was different from the segments used to form the reference pattern for that individual. 12-dimensional linear prediction (LP) coefficients were used to represent the speech segments of the speakers. For each segment, the unknown vector was correlated with the reference vectors and the correlations were averaged over a number of segments - the speaker with the largest correlation was identified as the unknown speaker.

The performance of text-independent systems was generally observed to be inferior to that of the text-dependent systems, which had two principal repercussions: (i) the pursuit of features other than the spectrogram, and (ii) the search for better time alignment techniques for matching training and testing utterances of identical phrases (Sambur 1973; Li and Hughes 1974; Atal 1974; Rosenberg and Sambur 1975; Beek et al. 1977; Markel et al. 1977; Furui 1981). Sambur, for example, in the Bell Labs, examined the speaker recognition and verification effectiveness of a set of 92 measurements, derived using LP analysis (Sambur 1973). These measurements included the formant structure of vowels, the duration of certain speech events, the dynamic behaviour of the formant contours, various aspects of the pitch contour throughout an utterance, formant bandwidths, glottal source 'poles', and pole and zero locations during the production of nasals and strident consonants. The experimental speech data were collected during five different recording sessions, distributed over three years. The measurements that were found most useful were related to the nasals, certain vowel resonances, certain temporal attributes, and average fundamental frequency. Another notable work, by Markel at. al., explored the effectiveness of long-term feature averaging for fundamental frequency-related, gain-related, and spectrally-related parameters (Markel et al. 1977).

Again, following Sambur's work, Atal, in the Bell labs, investigated the effectiveness of several different parametric representations obtained from LP analysis of speech, such as the predictor coefficients, the impulse response function corresponding to the transfer function based on the predictor coefficients, the autocorrelation function of the impulse response, the area function of an acoustic tube with an identical transfer function, and the cepstrum representing the logarithmic transfer function (Atal 1974). He tested these parameters for both text-dependent and text-independent speaker recognition (identification and verification), using a dataset of 60 utterances, consisting of six repetitions of the same sentence spoken by 10 speakers. Following were the highlights of Atal's work:

- The various parameters did not differ in speaker recognition performance by a wide margin; the cepstrum produced the best performance.
- The identification accuracy increased with the duration of the spoken material.
- The identification accuracy was not affected significantly by the removal of the time averages from the cepstrum samples.

Following Atal's work, Furui presented (published in 1981) a new set of techniques for a text-dependent speaker verification system for telephone-quality speech (Furui 1981). He introduced a new set of cepstrum-based features and a new method of overall distance computation, which revolutionised the research domain. The block-diagram of his system is shown in Fig. 1 Following were the highlights of his work:

- The time functions of the cepstrum coefficients were expanded by an orthogonal polynomial representation over short time segments. The first three orthogonal polynomials was used, producing coefficients which represent the mean value, slope, and curvature of the time function of each cepstrum coefficient in each segment, respectively.
- A sample utterance was brought into time registration with the reference template to calculate the distance between them. This was accomplished by a new time-warping method using dynamic programming technique, which came to be popularly known as dynamic time warping (DTW).
- As there is often some uncertainty in the location of both the initial and final frames due to breath noise, etc., the unconstrained endpoint technique was applied.

Starting with the 1980s, much more rapid and significant technological advancements were observed compared to the two decades prior to it, as research in speech recognition spilled over to speaker recognition, as discussed in the following sections.

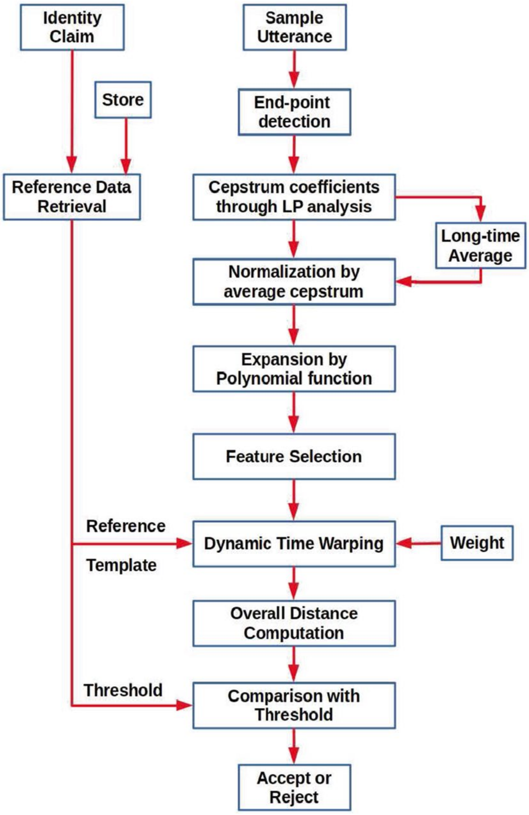


**Fig. 1** Block diagram of the speaker-verification system (Furui 1981)

## 3 Speaker recognition based on vector quantization and hidden markov models: 1980s

The inferior performance of text-independent speaker recognition compared to text-dependent speaker recognition not only led to the exploration of better acoustic features, as discussed previously, but also other techniques. Some researchers who believed that the context or timing information is irrelevant for speaker recognition, particularly in the text-independent scenario, explored Vector Quantization or the K-means clustering algorithm. In one such exploration, Soong et. al. used a 100-talker (50 male and 50 female) telephone recording database consisting of isolated digit utterances (Soong et al. 1987). A vector quantization (VQ) codebook was used as an efficient means of characterizing the short-time spectral features of a speaker, using a minimum distance (distortion) classification rule. The primary conclusions of this work were:

- Both larger codebook size and longer test token length (more digits in the test utterance) can be used to improve the recognition performance.
- VQ codebook should be updated from time to time to alleviate the performance degradation due to different recording conditions and intra-speaker variations.

Based on this work, Rosenberg and Soong modified the process to cater to text-dependent speaker recognition (Rosenberg and Soong 1987). The system, shown in Fig. 2, was evaluated using a 100-talker database of 20,000 spoken digits. Word prototypes for the text-dependent operation were constructed using the first 50 training utterances. The distance

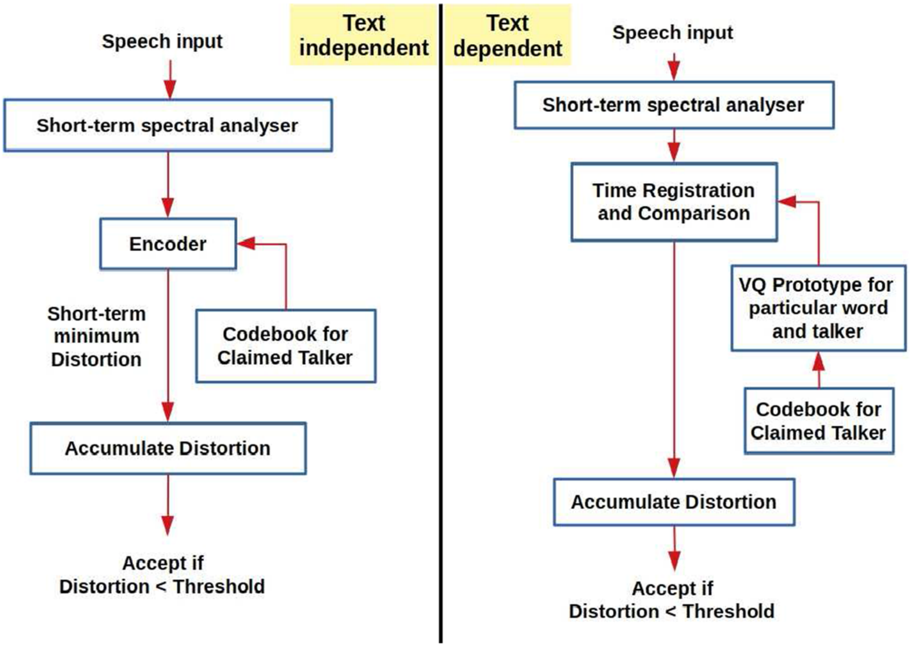

Text independent
Text dependent
Speech input
Short-term spectral analyser
Encoder
Short-term minimum Distortion
Codebook for Claimed Talker
Accumulate Distortion
Accept if Distortion < Threshold
Speech input
Short-term spectral analyser
Time Registration and Comparison
VQ Prototype for particular word and talker
Codebook for Claimed Talker
Accumulate Distortion
Accept if Distortion < Threshold

**Fig. 2** Block diagram of the VQ-based text-independent and text-dependent speaker verification process (Rosenberg and Soong 1987)

calculation was similar to the text-independent scenario, except that the distortion measurements are carried out with respect to codebook vectors specified by a word prototype instead of the best matching codebook vector. This system performed quite well and competed with the DTW method.

Meanwhile, inspired by the success of hidden Markov models (HMMs) in speech recognition, some researchers started exploring HMMs in speaker recognition. One of the first works in this direction was in the domain of text-independent speaker recognition by Poritz (Poritz 1982). In his experiments, Poritz used a set of 10 speakers uttering the same English passage. The chief ideas behind Poritz's work were:

- A speech signal is decomposed into a sequence of short segments of length $M$.
- Each segment is represented as the output of one of $S$ previously selected all pole recursive filters' sources. The filters are defined by polynomials of some degree, $N$, with $N < M$.
- The order in which the filters appear in the sequence is controlled by a previously selected state Markov chain.

Poritz's work was explored and improved upon by other researchers, most notably by Tishby (Tisby 1991). Tisby used a database of 20,000 isolated digit utterances, spoken by 100 speakers, 50 male and 50 female, over dialed-up local telephone lines. The main set of experiments used an automatically trained, VQ-based initial model with 8 states and 2, 4, and 8 mixtures in each state. In these auto-regressive (AR) or linear predictive HMMs, the states are described by a linear combination (mixture) of AR sources. In the case of an AR-HMM, the output/observation probabilities are determined by modelling the speech samples as a Gaussian autoregressive process. The important features considered were the spectral envelope parameters, or the related LPC or the LPC-derived cepstral coefficients. The system is shown in Fig. 3. Tisby concluded that the improvement in the performance of AR-HMMs over simpler techniques such as vector-quantization (VQ) was not substantially large to justify the effort in training them for speaker recognition.

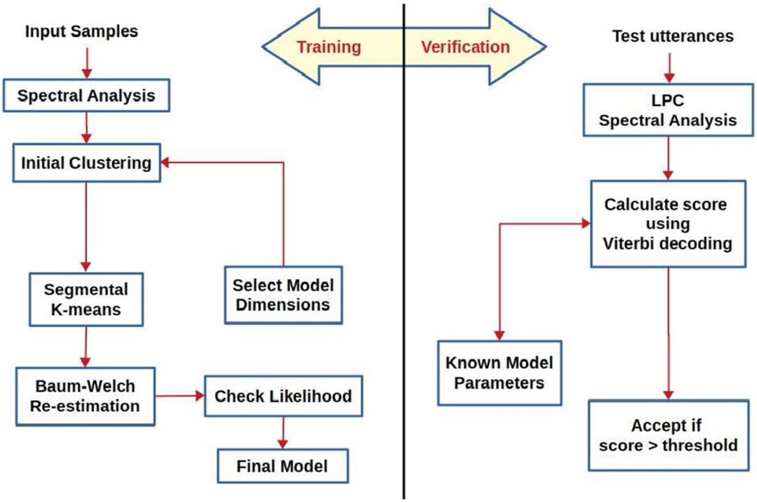


**Fig. 3** Block diagram of the HMM-based training and verification process (Tisby 1991)

Just as in text-independent speaker recognition, researchers also explored HMMs for text-dependent speaker recognition to test its efficacy against time-alignment techniques such as DTW. One of the notable works was conducted by Naik et. al., who collected two large speech databases for developing and analysing new speaker verification algorithms (Naik et al. 1989). The Handset database was collected in a controlled, laboratory setup with the aim of capturing variabilities across various types of telephone handsets in common use. Ten different handsets were used to gather the speech data. There were 20 speakers (10 men and 10 women). Five sessions of data were collected over a period of two weeks, with a total of 680 utterances for each of the 20 speakers. The Prototype database was collected over the long-distance public telephone network using a population of 100 speakers (65 Men, 35 Women). A total of 40 sessions for each speaker were collected over a period of four months. The salient points of this work were:

- Word-level HMM algorithms gave significantly better performance than whole-phrase DTW algorithms.
- A speaker discrimination model was used, which applied a linear transformation matrix to maximize the separability between true speakers and impostors in a given database. Large improvements in speaker verification performance were obtained from this speaker discrimination modelling.

As electronics and computing technology became faster and more affordable in the 1990 s, research in speaker recognition became much more diversified, which is discussed in the subsequent sections.

## 4 Text-independent speaker recognition based on Gaussian mixture models: 1990s

The findings of researchers like Tishby and Soong inspired a revised method of Speaker Recognition, one that combines the merits of VQ and HMM. This led to the development of the Gaussian Mixture Model (GMM), or the single-state HMM. Rose and Reynolds published the first work using GMM, based on experiments using a 12 reference speaker population (8 males, 4 females) from a conversational speech database (Rose and Reynolds 1990). The following were the key facets of this work:

- 20th order Mel frequency cepstral coefficients (MFCCs), with the 0th term removed, were used as features. This was one of the first works which utilized MFCCs as features for speaker recognition, almost a decade after they were proposed for speech recognition (Davis and Mermelstein 1980).
- GMM was found to share the robustness associated with the parametric Gaussian density function, while at the same time sharing the ability of non-parametric models to model non-Gaussian distributed data.
- Mixture modelling was shown to outperform the standard unimodal Gaussian classifier for all test utterance lengths.

Rose and Reynold's work was further supplemented and verified by Matsui and Furui, who made a thorough comparison of the VQ distortion-based speaker recognition method and discrete/continuous ergodic HMM-based ones, especially from the viewpoint of robustness against utterance variations (Matsui and Furui 1994). A database consisting of sentence data uttered at three speeds (normal, fast, and slow) by 23 male and 13 female talkers, recorded in three sessions over six months, was used for the experiments. Cepstral coefficients derived from LP analysis were used as features. Only speech spoken at normal speed was used for training, whereas all three types of speech were used for testing. The key points of this work were:

- A continuous ergodic HMM was found to be as robust as a VQ-distortion method when enough data is available, and a continuous ergodic HMM was far superior to a discrete ergodic HMM.
- Information on transitions between different states was found to be ineffective for text-independent speaker recognition. The speaker-identification rate using a continuous ergodic HMM was found to be strongly correlated with the total number of mixtures irrespective of the number of states.

## 5 Text-prompted speaker-verification: 1990s

One of the principal demerits of text-dependent speaker verification systems is that it uses a small set of pre-determined phrases, thus increasing vulnerability to determined imposters. As a counter to this problem, Higgins et al. proposed a system which could be considered as an intermediate between text-dependent and text-independent speaker verification systems (Higgins et al. 1991). The proposed system used a prompting strategy in which phrases are composed at random using a small vocabulary of words. The speech data consisted of ‘combination-lock’ phrases. Each phrase consisted of three numbers between 21 and 97. A verification trial or session consisted of four such phrases. Enrollment required speaking 24 such phrases, about 3 min per enrollment session. The spoken phrases were compared with word templates derived from enrollment sessions. Expectedly, whenever words occurred in the test material in contexts that did not occur in the enrollment material, the unmodeled coarticulations contributed to the high dissimilarity between the input speech and the claimant’s word templates, increasing the likelihood of rejecting valid users. As a counter to this problem, utterances were evaluated using the claimant’s model and using an alternate model composed of templates from other speakers, called the ‘cohort speakers’ or ‘ratio-set’, as shown in Fig. 4. Since all the speakers were equally affected by the unknown context, the likelihood ratio score was used as a robust metric. The likelihood ratio score represents approximately the ratio of the probability of the observed input assuming it was spoken by the claimant to the probability of the observed input assuming it was spoken by the cohort speakers. This was one of the first works to utilize the concept of ‘cohort speakers’ in speaker recognition, and the first to use the technique of ‘likelihood ratio scoring’.

Following the work of Higgins, Furui et. al. investigated two methods that captured both the phoneme and speaker information from a small amount of training data for a text-varying speaker verification system (Matsui and Furui 1993). Two methods of making speaker-specific phoneme models were experimented with, using a database of 15 speakers (10 male, 5 female) recorded in three sessions over six months. Method-I used phoneme-adaptation of a phoneme-independent speaker model, whereas Method-II is based on speaker-adaptation of universal phoneme models. Cepstral coefficients derived using LP analysis were used as features in this work. For speaker verification, Method-I was found to be more

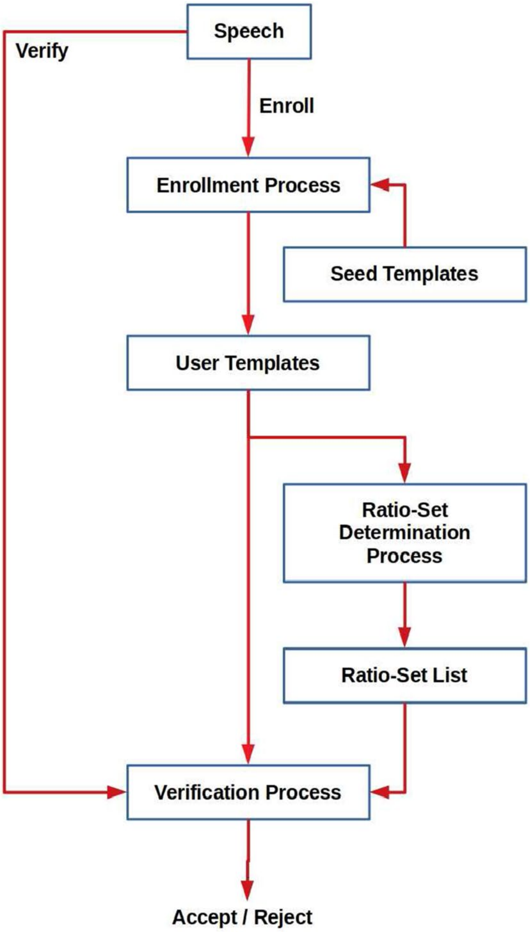


**Fig. 4** Block diagram of the speaker verification system using randomized phrase prompting (Higgins et al. 1991)

efficient than Method-II, and for rejection of incorrect speech, the reverse was true. It was also found that either of these methods when supplemented with a phoneme-independent speaker model (to make up for the lack of speaker information), improved the speaker verification performance.

Unfortunately, text-prompted speaker-verification systems could not perform at par with fixed-phrase speaker-verification systems, particularly when the number of speakers was large.

## 6 Score normalization for speaker recognition: 1990s and 2000s

Speech being a highly non-stationary signal, it varies even when it is spoken by the same person at different times. These differences are further accentuated by the channel variations, and miscellaneous factors such as the health, emotion, mood, etc. of the speaker. As such, determining a universal threshold score at which a speaker could be accepted or rejected in a speaker verification system is a challenge. In Higgins' work, discussed previously, a ‘likelihood ratio score’ was proposed, which used the concept of ‘cohort speakers’ (Higgins et al. 1991). Matsui and Furui realised that one key issue in Higgins’ work was the appropriate selection of cohort speakers. Hence, they proposed an alternative method in which the summation of the similarity values of the models of the registered speakers was approximated by the similarity value of a pooled model (Matsui and Furui 1994). Two methods of pooling were used: one using a separate population of speakers, and the other using all the registered speakers. The pooled model was created using HMM. This is one of the first works which utilized the concept of a ‘universal background model’ or UBM, consisting of a separate population of speakers. The experiments were conducted on a sentence database of 30 speakers (20 male, 10 female) recorded in five sessions over ten months. Cepstral coefficients based on LP analysis were used as features. Speaker verification (without using cohort speakers) generally required comparing each speaker with itself and the rest of the speakers as imposters. The proposed methods significantly reduced the amount of calculation needed for normalisation and performed the same or better than the standard methods. However, the speaker verification performance (for both text-independent and text-prompted) was observed to be poorer when different populations of customers and imposters were used.

Following Furui’s work, Rose and Reynolds published their famous seminal work using GMM for speaker identification and verification with background speaker normalization and a likelihood ratio test (Reynolds 1995). This was the completion of their introductory work published in the early 90s. A novel technique for selecting ‘background speakers’ (similar to ‘cohort speakers’ but not the same) was also presented. The block diagrams of the system are showcased in Fig. 5. The systems were evaluated on four publicly available large speech databases of the time: TIMIT (630 speakers), NTIMIT (630 speakers), Switchboard (500 speakers) and YOHO (138 speakers). Each of these databases possessed different characteristics both in task domain (e.g., text-dependency, number of speakers)

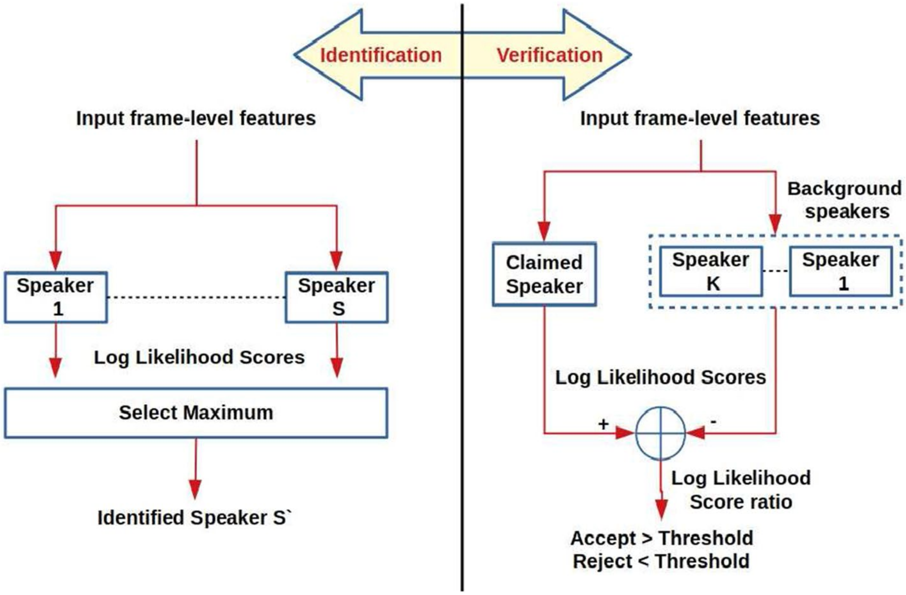


**Fig. 5** Block diagram of the speaker identification and verification system (Reynolds 1995)

and speech quality (e.g., clean, wideband, noisy telephone) allowing for experimentation over a wide variety of tasks and conditions. Rose and Reynolds also promoted the use of Mel-scale cepstral coefficients (MFCCs) as frame-level features over other features.

In 2004, Bimbot et. al. published a review article on text-independent speaker verification, which included many normalization methods used in different scenarios, such as world-model and cohort-based normalization, centred/reduced impostor distribution normalization, Znorm, Hnorm, Tnorm, H-Tnorm, Cnorm, Dnorm, and WMAP (Bimbot et al. 2004). Out of these, the most commonly used normalization is the centred/reduced impostor distribution normalization wherein the scores are normalized by subtracting the mean and then dividing by the standard deviation, both estimated from the (pseudo) impostor score distribution.

## 7 Focus on real-world speaker recognition: 2000s

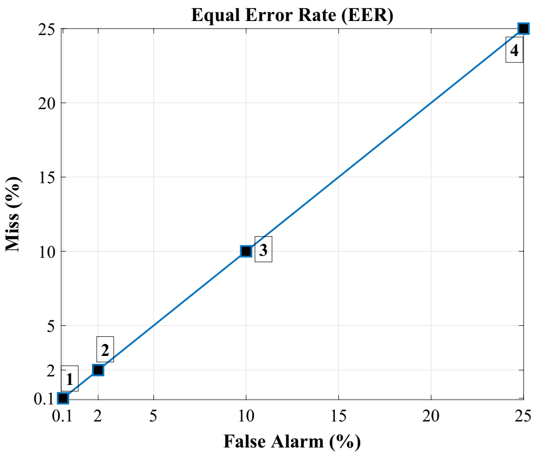
As the new millennium ushered in, many prominent researchers expressed their views on the current state and the future of speaker recognition technology. An important assessment was made by Reynolds who evaluated the performance of four different speaker verification systems (Reynolds 2002):

1 Text-dependent using combinations lock phrases (e.g., 35-41-89): Clean data was recorded using a single handset over multiple sessions, with ~3 min of training data and ~2 s test data.
2 Text-dependent using ten-digit strings: Telephone data recorded using multiple handsets with multiple sessions was used. Two strings were used for training and a single string for testing.
3 Text-independent using conversational speech: Telephone data using multiple handsets with multiple sessions. Two minutes were used for training and 30 s for testing.
4 Text-independent using read sentences: Very noisy radio data was collected using multiple military radios and microphones with multiple sessions. 30 s was used for training and 15 s for testing.

The performance of the four systems are shown in Fig. 6. In the figure, equal error rate (EER) represents the error rate at which both the miss (genuine claim not verified) and false alarm (imposter claim verified) are equal. At the particular EER, the decision threshold provides the same miss rate and false alarm rate. The EER, thus, provides an unbiased metric of the system performance. Comparing the EERs, Reynolds concluded that robustness to channel variability was the biggest challenge to speaker recognition, and outlined the following key focus areas:

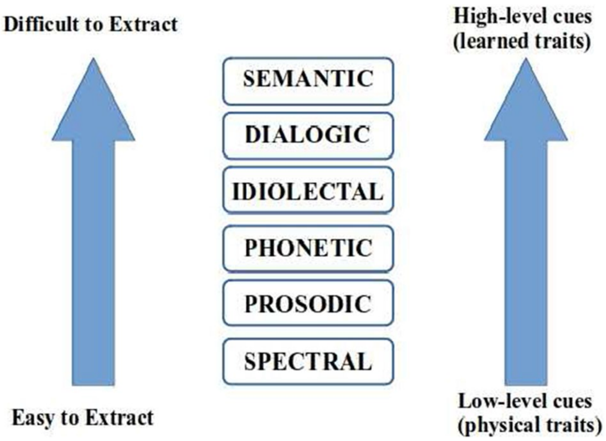
- Exploitation of higher-levels of information which are more robust than low-level spectrum features.
- Obtaining speech from a wide variety of handsets, channels and acoustic environments that would allow examination of problem cases and development and application of new or improved compensation techniques.
- Emphasis on unconstrained tasks, such as variable channels and noise conditions, text-independent speech and the tasks of speaker segmentation and indexing of multispeaker speech.

**Fig. 6** Performance of the four speaker verification systems (Reynolds 2002)



Similar observations were made by other researchers, such as Zanuy et al., who, specifically, emphasized the need to explore higher-level features (Fig. 7) and the fusion of both learned and physical traits (Faundez-Zanuy and Monte-Moreno 2005). It was also widely recognized that out of the two branches of speaker recognition (speaker identification and speaker verification), speaker verification was the most viable for commercial and forensic purposes as an added biometric or security layer. Text-independent speaker verification added more flexibility and was more 'hack-proof', and hence was envisaged as the future of 'voice-print' technology. In fact, Zanuy et al. noted that "unlike other biometric technologies, with voice input, the system has other degrees of freedom, such as the use of knowledge/codes that only the user knows, or dialectical/semantical traits that are difficult to forge".

**Fig. 7** Levels of information for speaker recognition (Faundez-Zanuy and Monte-Moreno 2005)



## 8 Feature engineering for speaker recognition: 2000s

As can be inferred from the previous section, in the first decade of the new millennium, a lot of emphasis was given to exploring features which represented discriminatory speaker-specific cues. The short-term acoustic features, such as the MFCCs, and various features derived from LP-analysis still remained the widely used and accepted features for speaker recognition (Kinnunen and Li 2010). Some researchers, such as Lu and Dang, tried to challenge this traditional approach of feature engineering (Lu and Dang 2008). In their study, the NTT-VR speaker recognition database was used, consisting of 35 speakers (22 male and 13 female). The speech was collected in five sessions over a period of 10 months. In each session, each speaker was asked to speak sentences with normal, slow and fast speech rates. The average duration of the sentences was about 4 s. The following were the highlights of this work:

- The speech spectrum was divided into uniform overlapping bands, constituting a uniform filterbank. The F-ratio, quantifying the inter-speaker variance vs. the intra-speaker variance, was calculated for each frequency band. Similarly, mutual information scores quantifying discriminatory speaker information in each frequency band were calculated.
- The plots of F-ratio vs. frequency-band and mutual information vs. frequency-band showed that speaker discriminatory information is predominantly present in the low-frequency and the high-frequency regions of the spectrum, and is the least in the mid-frequency region. This result challenged the use of the Mel filterbank which steadily de-emphasizes the higher frequency spectrum and puts the most emphasis on the low-frequency spectrum.
- A non-uniform filterbank was constructed based on the F-ratio vs. frequency curve, and then used to extract features in the same manner as the MFCCs. These features were found to perform better than the MFCCs in speaker identification for the NTT-VR database.

While some researchers were in pursuit of completely new ways of feature engineering, others were in the hunt for features which could be fused with traditional features in a complementary manner. Spectro-temporal features and prosodic features fall in this category (Kinnunen and Li 2010). The most widely used spectro-temporal features, even today and at that time, are the first and second time-derivatives of the MFCCs, called the velocity (or $\Delta$) and acceleration (or $\Delta\Delta$) coefficients, which are appended to the MFCCs to construct an enlarged feature-vector per speech frame. Possibly inspired by the success of the $\Delta$ and $\Delta\Delta$ coefficients, some researchers explored other possibilities. One such exploration was done by Kinnunen et al., who proposed the 'modulation spectrum features' (Kinnunen et al. 2008). The modulation spectrum features basically represented the average magnitude spectrum across a continuous span of speech frames. The test scores derived from these features were merged with the test scores obtained from the MFCCs, resulting in better speaker recognition performance.

Unlike the spectro-temporal features, prosodic features explicitly try to capture the idiosyncrasies of the speaker such as his/her speaking rate, duration of phones, energy distribution, pitch patterns and other voice-source features. In spite of increased complexity (in comparison with the spectro-temporal features), a plethora of research was devoted to extracting prosodic features during the 2000 s. One notable work was contributed by

Shriberg et al., who made a thorough investigation using pitch, duration and energy features (Shriberg et al. 2005). In this work, any given speech utterance was first transcribed by a speech recognition engine, which was then re-transcribed as syllables. Phone-level alignment information was obtained from the speech recognizer, and then various features related to the duration, pitch, and energy values in the syllable were computed. It was observed that pitch level or fundamental frequency was the most useful marker of speaker identity. Prosodic features, however, are not as easily obtained as spectro-temporal features. For example, obtaining features related to the voice-source requires access to the glottal excitation signal which is buried in the non-stationary speech signal. For speech recorded in a noisy environment, the features or parameters estimated are not so reliable. As a result, many researchers continue to hunt for the magic algorithm.

One of the popular and reliable sources for extracting voice-source features is the residual signal obtained from the LP analysis of speech. As such, some researchers, such as Prasanna et al., used the residual signal to extract features which could complement the state-of-the-art speaker recognition systems, but without the added complexity of a speech recognition engine (Prasanna et al. 2006). The system is shown in Fig. 8. In this work, a small database of 20 speakers was used for speaker identification, and the much larger 'National Institute of Standards and Technology Speaker Recognition Evaluation' (NIST SRE) 2002 database was used for speaker verification. The NIST SRE 2002 database consisted of approximately 157 h of speech-data collected from different regions (roughly around 500 speakers per region) of the USA. The highlights of the work were:

- Important voice-source information is hypothesized to be contained in the non-linear and 'higher order' relations among the samples of the LP residual, which represents the excitation signal according to the source-filter theory of speech production.
- To capture the relationship among the samples of the LP residual of any voiced-speech frame (frame corresponding to vowel or vowel-like sounds), an auto-associative neural network (AANN) is used, in which the input vector and the output labels correspond to the LP residual samples. The error between the output layer and output labels is represented in terms of a confidence score.

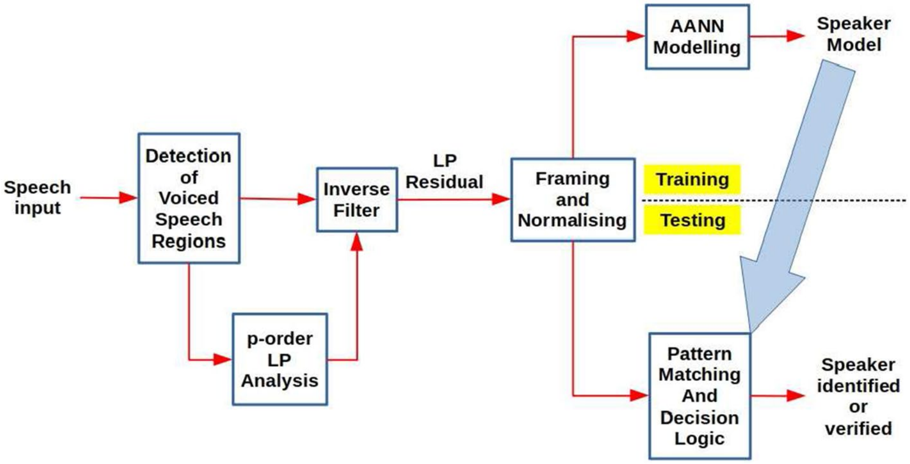


**Fig. 8** LP-residual for speaker recognition using AANN modelling (Prasanna et al. 2006)

- An AANN is trained for the voiced-speech frames of each speaker, and the confidence score is used for testing against any random speech utterance. The confidence scores were merged with the likelihood scores obtained from GMM-UBM systems trained with standard acoustic (spectral) features to obtain significant performance improvement.

As shown in Fig. 7, even higher levels of speaker information could be derived provided sufficient data and computing power is available. This search beyond prosodic features was pioneered by Doddington (Doddington 2001). The objective of his famous work was to encapsulate the speaking habits of speakers, reflected in their long-term speech patterns, such as the usage of certain words and phrases. Doddington investigated the use of $N-$gram word models to encapsulate the 'idiolectal' differences of speakers. An $N-$gram model, simply put, provides the probability of occurrence of a given sequence of $N$ words in the vocabulary. Doddington found bi-gram and tri-gram models to be the most effective, though their performance could not be compared with acoustic spectral features. Inspired by Doddington's work, many pursued research in the exploration of higher-level speaker-specific information. For example, Ma et al. used multiple GMM-based 'tokenizers' in parallel, different tokenizers representing different levels of speaker information - low to high (Ma et al. 2006). Such works were not purely the brainchild of research in speaker recognition but were inspired by other domains of speech processing, mainly language recognition and speech recognition. Nevertheless, extracting higher-level information does require a significant amount of data, which is not user-friendly, particularly during testing.

## 9 Refined and alternative models for speaker recognition: 2000s

Along with features, there was a concurrent exploration of methods to make the models used for speaker recognition more robust. These explorations led to a variety of modelling techniques, which could be broadly identified as the refinement of the existing GMM-based systems, exploration of discriminative modelling techniques, and fusion of different modelling systems using the same or different features (Kinnunen and Li 2010). Following their seminal work, Reynolds et al. proposed their equally famous maximum a posteriori (MAP) adapted GMM system (Reynolds et al. 2000). In this approach, a set of background speakers are used to create a speaker-independent universal background model (UBM) using the vanilla GMM. The speaker-specific models are created by adapting the GMM-UBM parameters, more specifically the means of the GMM-UBM using Bayesian statistics. This process enabled more reliable models when there was limited data available, particularly in unseen channel conditions. As such, the MAP-based GMM became the baseline for speaker recognition in the 2000 s. Other techniques for adapting the GMM-UBM, such as maximum likelihood linear regression (MLLR), were also explored (Kinnunen and Li 2010). Further, it is observed that the design of speaker recognition is limited to the domain variations (including, recording device, environment, etc.), as the GMM-UBM models the speaker along with the environment acoustics. To address the issue Kenny et al. in (Kenny 2005; Kenny et al. 2008), proposed a vector space-based approach called joint factor analysis (JFA) for speaker verification. The approach provides a significant improvement in speaker verification performance over the GMMs, by modelling the inter-speaker variabilities. The key aspects of JFA are as follows:

- Consider a GMM-UBM of $C$ mixtures, derived from $F$-dimensional acoustic feature-vectors. A supervector, $\mathbf{m}$, of $CF$-dimension is constructed by concatenating the $F$-dimensional means of the GMM-UBM.
- Given a training feature-set for a given speaker, the speaker-dependent and channel-dependent supervector, $\mathbf{m}_s$, is assumed to be of the form,

$$\mathbf{m}_s = \mathbf{s} + \mathbf{c} = (\mathbf{m} + \mathbf{V}\mathbf{y} + \mathbf{D}\mathbf{z}) + \mathbf{U}\mathbf{x}. \tag{1}$$

  In the above, $\mathbf{s}$ and $\mathbf{c}$ represent the speaker and channel dependent components, respectively; $\mathbf{V}$ is a rectangular matrix whose columns are called eigenvoices; the elements of $\mathbf{y}$ are called speaker-factors; $\mathbf{D}$ is a square diagonal matrix; $\mathbf{z}$ captures the session-variability; $\mathbf{U}$ is a rectangular matrix and $\mathbf{x}$ represents the channel-factors.
- The matrices $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{D}$ are estimated a priori on large datasets: $\mathbf{V}$ followed by $\mathbf{U}$ and $\mathbf{D}$. For a given training sample, $\mathbf{x}$ and $\mathbf{y}$ are first estimated jointly, and then $\mathbf{x}$. Lastly, $\mathbf{c}$ is discarded and $\mathbf{s}$ is used as the speaker model.

The performance of JFA was evaluated on the NIST SRE 2006 dataset, consisting of over 400 h of conversational telephone and microphone speech data, and included speech in languages other than English. Soon, the JFA system came to be recognized as one of the top-performing speaker recognition systems, comfortably ahead of the MAP-adapted GMM system.

While GMM-based systems were being developed, in parallel, there were significant developments in utilizing discriminative modelling techniques, such as support vector machines (SVMs) and artificial neural networks (Kinnunen and Li 2010). Generative models like the GMM represent the feature-space of the speaker in terms of a pre-selected statistical distribution. For the GMM, that distribution happens to be a weighted combination of multi-dimensional normal distributions. Discriminative models such as SVMs, however, model the boundary between the feature-space of the speaker and (ideally) that of the rest of the population. There were many significant works trying to use discriminative modelling as a competitor to (or complement) the GMM-based speaker recognition systems. For example, recall Prasanna's work using AANNs discussed in the previous section (Prasanna et al. 2006). Nevertheless, amongst the multitude of noteworthy works, the method proposed by Campbell et al. using SVMs stood out (Campbell et al. 2006). The NIST SRE 2003 database, which consisted of over 100 h of conversational speech data, was used in Campbell's work. The highlights of the work were:

- The SVM is a binary classifier which transforms the features to a higher dimension using a Kernal function prior to estimating the class boundary.

$$\begin{aligned} &f(\mathbf{x}) = \sum_i \alpha_i t_i K(\mathbf{x}, \mathbf{x}_i) + d, \\ &t_i = \{-1, 1\};\ \sum_i \alpha_i t_i = 1, \alpha_i > 0; \\ &K(\mathbf{x}, \mathbf{x}_i) = [\mathbf{b}(\mathbf{x})]^{\mathrm{T}}[\mathbf{b}(\mathbf{x}_i)]. \end{aligned} \tag{2}$$

  In the above equation-set, $t_i$ represents the ground truth or output labels, which in the case of speaker recognition simply indicates whether or not the features belong to the target speaker. $K(., .)$ represents the Kernal function, which uses a mapping function, $\mathbf{b}(.)$, to transform a given feature vector to a higher-dimensional space. The vectors, $\mathbf{x}_i$, are called the support vectors and are obtained from the training data by an

optimization process. The class decision is determined by comparing $f(\mathbf{x})$ to an appropriately determined threshold.
- A novel and computationally efficient sequence kernel was derived called the generalized linear discriminant sequence (GLDS) kernel.
- The GLDS-SVM performed comparably to the MAP-adapted GMM. Significant relative performance enhancement was observed when it was fused (at score level) with the MAP-adapted GMM.

The GLDS-SVM also performed competitively with other state-of-the-art systems on the NIST SRE 2006 database which included speech data of multiple languages (Kinnunen and Li 2010).

Apart from the development of individual speaker recognition models competing for supremacy, many-a-research also focused on creating hybrid systems, combining multiple different modelling techniques (Kinnunen and Li 2010). Campbell et al., for example, also proposed a GMM-SVM model which used supervectors obtained from MAP-adapted GMMs as input features to a SVM (Campbell et al. 2006). Brummer et al. proposed the widely appreciated STBU system, which was a combination of three different subsystems (Brummer et al. 2007):

(1) GMM, with MFCCs or perceptual linear prediction (PLP) features.
(2) GMM-SVM.
(3) Maximum-likelihood linear regression support vector machine (MLLR-SVM), using MLLR speaker adaptation coefficients derived from an English large vocabulary continuous speech recognition (LVCSR) system.

Many other notable fusion systems were explored. These fusion systems, generally, performed better than the individual systems at the expense of overall complexity. Such research explorations, however, provided insights into the complementary or redundant nature of different types of modelling techniques, and the optimal features for them.

## 10 Modern speaker recognition: 2010s and present

The continued advancements in computing technology and the unabated search for solutions for the key challenges identified in the previous decade resulted in further diversification of research groups in speaker recognition.

The search for complementary features continued. In one such notable work, Nakagawa et al. proposed the combination of phase information with the MFCC features, in the context of text-independent speaker verification (Nakagawa et al. 2011). The average normalized unwrapped phase values were computed for different sub-bands of the spectrum, to train a phase-GMM. The phase values were mapped to the coordinates of a unit circle. The weighted average of the likelihoods obtained for test utterances was then computed for speaker identification. Although the bi-variate GMMs with 64 mixtures provided increased equal error rates (EERs), the score level combination outperformed the individual MFCC-GMM system for NTT and large-scale JNAS (Japanese News Article Sentences) databases. Other works, such as by Sharma et al., explored the use of complementary features derived from modern non-linear and non-stationary signal processing techniques (Sharma et al. 2017; Sharma et al. 2018). Sharma et al. reported that short-term features derived using

Empirical Mode Decomposition (EMD) of speech could be used, instead of the higher dimensions of the MFCCs and their Δ and ΔΔ coefficients. In (Sharma et al. 2017), the features derived from EMD were found to be more useful than the higher dimensions of the MFCCs, when tested on the NIST-SRE 2003 (normal speech) and the CHAINS corpus (normal, fast and whispered speech), for text-independent speaker verification. In (Sharma et al. 2018), a spectral representation derived from the EMD of speech, called the Hilbert spectrum, was used to extract features which were tested for text-dependent speaker verification. Results similar to that reported in (Sharma et al. 2017) was observed on the RSR2015 and IITG (Dey et al. 2014) databases, particularly under noisy testing conditions.

Krishnamoorthy et al. proposed an approach for improving the GMM and GMM-UBM speaker recognition performance by adding white noise at moderate levels (SNR of 20 dB and 40 dB) (Krishnamoorthy et al. 2011). The work was proposed under the condition that limited data is available for training and testing. The GMM and GMM-UBM performances on randomly selected 100 speakers data for TIMIT databases showed improvement when the speaker utterances were augmented by adding white noise. For the development of robust speaker verification systems, Prasanna et al. proposed extraction of MFCC features by segmenting the speech into vowel-like regions (VLR) (Prasanna and Pradhan 2011). An algorithm was proposed to estimate the VLRs by detecting the vowel onset points using zero-frequency filtering of speech. The MFCC features estimated from VLRs showed a significant reduction in the EER when evaluated on the NIST SRE 2003 database. The speaker verification performance was found to be unaffected even after adding white and factory noises at various SNR levels ranging from 9 dB (moderate) and 0 dB (extreme).

From the technology standpoint, the introduction of i-vector systems, as a combined refinement of the JFA and GLDS-SVM systems, dramatically improved the performance of the speaker verification systems in terms of reduced EER (Dehak et al. 2010; Kenny 2010). The JFA efficiently models the intra-speaker variability. However, it was observed in (Dehak et al. 2010) that the channel factor ($\mathbf{x}$) of JFA also carries speaker information. Alternatively, the GLDS-SVM approach efficiently models the speaker by compensating intra-speaker variability using nuisance attribute projection (NAP). Inspired by both techniques, the i-vector approach combined the speaker, session, and channel sub-space into a single subspace called the total variability space ($\mathbf{T}$), and the factor associated with this was called the i-vector($\mathbf{w}$). Further, the speaker representations were extracted by compensating the intra-speaker variability of the i-vectors. The key aspects of the i-vector system were as follows:

- The modified projection of the speaker-and-channel-dependent supervector was assumed to be of the form,

$$\mathbf{m}_s = \mathbf{m} + \mathbf{T}\mathbf{w}. \tag{3}$$

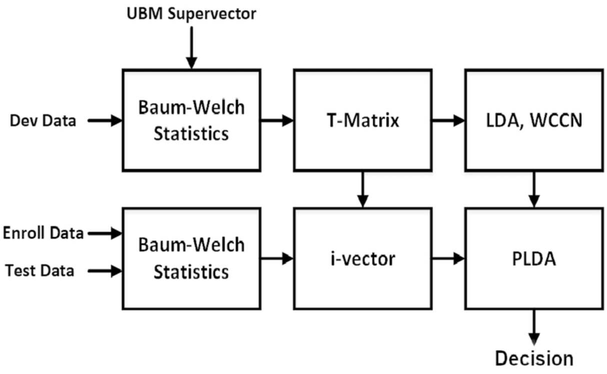
- Different channel compensation techniques: within class covariance normalization (WCCN), linear discrimination analysis (LDA), NAP, and their combinations were explored. The experiments concluded that the combination of LDA followed by WCCN outperforms the rest.
- SVM that used cosine similarity score, and the simple cosine similarity score, both, were used to compute the speaker similarity. It was observed that the direct use of cosine similarity was better. Later, Kenny et al. refined the speaker similarity computation approach by replacing cosine similarity with probabilistic linear discriminate analysis (PLDA) (Kenny 2010).

The block diagram of the i-vector speaker-verification system is depicted in Fig. 9. The success of the i-vector system inevitably resulted in a gamut of methods to refine the system (Haris et al. 2013; Pradhan and Prasanna 2013; George et al. 2018). Pradhan et al., for example, independently trained the i-vector speaker models for vowel-like regions (VLRs) and non-vowel-like regions. During verification, the scores obtained from both the i-vector models were combined by providing more weight to the speaker model corresponding to the VLRs. Improvement in the EER was observed when scores given by both models were combined by weighted averaging, with more weight to the VLR speaker models. Improvements in EER confirmed the presence of more speaker-specific information in the VLR regions compared to that of the non-VLR regions (Pradhan and Prasanna 2013). In a related work, Haris et al. proposed a speaker verification system by combining scores obtained from the i-vector based speaker verification system using VLR-based signal conditioning and a generic i-vector for microphone and telephone channel compensation (Haris et al. 2013; Haris 2014). Another work, by Senoussaoui et al. proposed a generic i-vector speaker verification system for channel compensation using the telephone microphone speech with the generic i-vector factorization as $\mathbf{s} = \mathbf{m} + [\mathbf{T}_{\text{phn}} + \mathbf{T}_{\text{mic}}]\mathbf{w}$. Here, $\mathbf{m}$ is the utterance-independent vector, $\mathbf{T}_{\text{phn}}$ and $\mathbf{T}_{\text{mic}}$ are the total variability matrices learnt from telephonic data and microphone data, respectively, and $\mathbf{w}$ is the speaker-specific i-vector (Senoussaoui et al. 2010). Later, speaker verification systems were proposed by exploiting the sparsity in the speaker utterances. In sparse representation coding (SRC) systems, the mean adapted vectors were represented as:

$$\mathbf{y} = \mathbf{D}\mathbf{x} \tag{4}$$

In Eq. 4, $\mathbf{D}$ is the learnt dictionary and $\mathbf{x}$ is the sparse coded vector. In SRC, the sparse representation of test utterances was obtained from a learnt exemplar dictionary for speaker verification. The JFA-based factorization was performed for channel/session compensation (Haris and Sinha 2011). The SRC approaches vary based on the methods used for learning the exemplar dictionary. The first SRC system, applied on the NIST SRE 2012 dataset, used the orthogonal matching pursuit (OMP) algorithm for extracting the sparse vector (Haris and Sinha 2011; Pati et al. 1993), whereas the second system used the K-singular value decomposition (KSVD) algorithm for learning the exemplar dictionary (Haris and Sinha 2012; Aharon et al. 2006). The final ‘NIST SRE 2012 IITG SV’ system used the speaker i-vector models obtained from generic i-vector system, VLR and non-VLR i-vector

**Fig. 9** Block diagram of the i-vector-based speaker-verification system (Dehak et al. 2010)



models and SRC systems with over-complete dictionaries learnt using OMP (trained using JFA Type-1 data) and KSVD (trained using JFA Type-2) algorithms.

The rapid rise and success of artificial intelligence drastically changed the landscape of speaker recognition, as with almost all areas of science and technology. The state-of-the-art techniques of speaker modelling which were mainly based on signal processing (applied mathematics) and conventional machine learning, came to be seriously challenged by deep learning technologies, which we discuss separately in the following sub-section.

### 10.1 Deep learning in speaker verification

As discussed in the previous section, i-vector systems with PLDA scoring were state of the art in the early 2010 s. During the process of developing fully fledged deep learning-based speaker verification systems, many researchers tried to replace the various components of an i-vector extractor (Stafylakis et al. 2012; Vasilakakis et al. 2013; Kenny et al. 2014; Gupta et al. 2014; Novotny et al. 2016). For example, in the work by Stafylakis et al., PLDA scoring in the backend of the i-vector extractor was replaced by a deep belief network to get similar performance (Stafylakis et al. 2012). A pseudo i-vector extractor was proposed using a stack of restricted Boltzmann machines which non-linearly transform the speech frames to a score predicted at the output unit. The distribution of the output score predicted by the network was hypothesised to be a pseudo i-vector. Pseudo i-vectors showed performance comparable to i-vector with LDA/cosine distance scoring but performed inferior to the i-vector PLDA system. Motivated by the effectiveness of deep neural networks (DNNs) in speech recognition, Kenny et al. used DNNs to capture the Baum Welch parameters (identical to zero and first-order Baum-Welsch statistics) of the UBM to replace the expectation maximization (EM) algorithm used in GMM-UBM (Kenny et al. 2014). The speaker verification systems which used DNN-based UBM provided 16% relative improvement in NIST SRE 2012 when scores were fused with traditional i-vector models with PLDA scoring. In another work, Chen et al. carried out frame-level discriminative learning of the deep neural networks for speaker information. The frame level embedding obtained from the network was fed to the GMM-based speaker models for verification (Chen and Salman 2011). Even though the aforementioned methods provided evidence of speaker information learnt by the networks, their performance was inferior to the baseline performance of the i-vector with PLDA scoring.

In 2014, Google introduced a text-dependent speaker verification system targeted for low-footprint devices, i.e., devices with limited computational resources (Variani et al. 2014). Deep vectors (d-vectors) obtained from the DNN were used as the speaker models during the speaker enrollment. During the evaluation phase, the scores were generated by computing the distance between speaker d-vectors and the test d-vectors. As illustrated in Fig. 10, d-vectors are computed as the average of the frame level embeddings obtained at the last hidden layer (excluding the final output layer of the feed-forward DNN). When more text-dependent utterances were used per speaker (20 utterances per speaker) during enrollment, the d-vector and i-vector systems showed comparable performance. The primary advantage of the d-vector system over the i-vector system was that no speaker adaptation was required during the enrollment making it suitable for low-footprint applications. As part of the ‘Ok Google’ project, the d-vector text-dependent speaker verification system was further extended towards the development of an end-to-end speaker verification system (Heigold et al. 2016). In the end-to-end system, the development, enrollment and evaluation stages of the speaker verification system were integrated into a single network, as

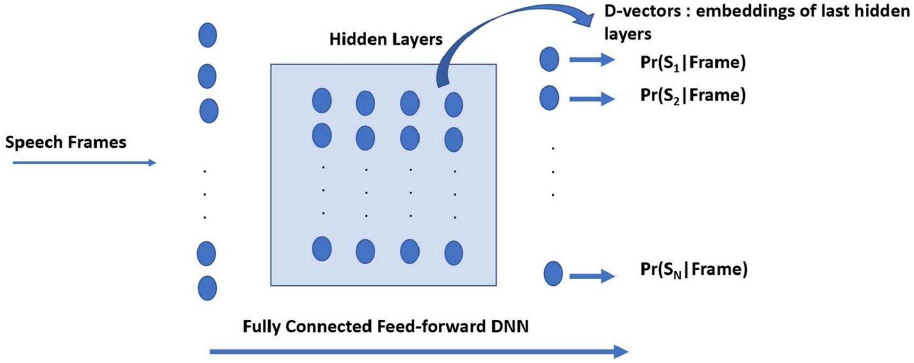
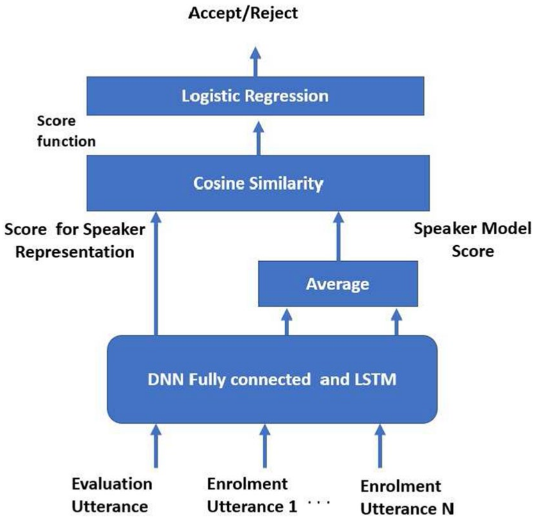

D-vectors : embeddings of last hidden layers
Hidden Layers
Pr($S_1$|Frame)
Pr($S_2$|Frame)
Speech Frames
Pr($S_N$|Frame)
Fully Connected Feed-forward DNN

**Fig. 10** Deep vectors for speaker modeling (Variani et al. 2014)

shown in Fig. 11. In the end-to-end network, the speaker representation score was generated from a fully connected DNN and speaker models of the enrolled speakers were generated from long short-term memory (LSTM) recurrent neural networks. As compared to the speaker representation obtained from d-vectors, in the end-to-end network, DNNs used a stack of longer fixed-length windows to generate the score corresponding to the speaker representation. A logistic regression function was then used for computing the probability of acceptance for verification.

In 2016, in the context of text-independent speaker verification, Snyder et al. proposed an end-end deep architecture for generating the speaker embeddings which they named as x-vector speaker embeddings (Snyder et al. 2016). A deep architecture based on network-in-network (NIN) layers was proposed to generate the speaker embeddings. This network trains a speaker model by learning the statistics over a stack of MFCC features extracted from a given utterance. During training, the network tries to optimize the probability of the embedding (**x**) of the utterance that belongs to the same speaker and minimizes the embedding (**y**) that belongs to the other speaker model.

**Fig. 11** End-to-end text-dependent speaker verification network for enrollment and verification (Heigold et al. 2016)



$$P(\mathbf{x},\mathbf{y}) = \frac{1}{1+e^{-L(\mathbf{x},\mathbf{y})}}, \tag{5}$$

$$L(\mathbf{x},\mathbf{y}) = \mathbf{x}^{\mathrm{T}}\mathbf{y} - \mathbf{x}^{\mathrm{T}}\mathbf{S}\mathbf{x} - \mathbf{y}^{\mathrm{T}}\mathbf{S}\mathbf{y} + b. \tag{6}$$

$$\begin{aligned} E = & - \sum_{\mathbf{x},\mathbf{y}\epsilon P_{\mathrm{same}}} \ln\{P(\mathbf{x},\mathbf{y})\} \\ & - K \sum_{\mathbf{x},\mathbf{y}\epsilon P_{\mathrm{diff}}} \ln\{1 - P(\mathbf{x},\mathbf{y})\}. \end{aligned} \tag{7}$$

Fig. 12 shows the deep learning architecture used for generating speaker embeddings from features. The probability of the two embeddings, **x** and **y**, belonging to the same speaker is computed by logistic regression and is given by Eq. 5. The cost function $L(\mathbf{x},\mathbf{y})$ is computed based on Eq. 6, where **S** is a symmetric matrix similar to what is obtained during singular value decomposition (SVD) (Strang 2009) and $b$ is the bias. The objective function of the optimization problem is given by Eq. 7, where $K$ is a constant scale factor that corresponds to the weight used for different speaker pairs. The probability of the two embeddings belonging to the same speaker is given by $P_{\mathrm{same}} = P(\mathbf{x},\mathbf{y})$ and that to different speakers is given by $P_{\mathrm{diff}}$=1 $- P(\mathbf{x},\mathbf{y})$. The enrol and test utterance scores are computed based on Eq. 6. The x-vector systems for end-to-end text-independent speaker-verification systems provided 13% relative improvement in comparison to the i-vector speaker-verification systems with PLDA backend.

In 2017, Synder et al. further refined the end-to-end network-based speaker verification system (Snyder et al. 2017). The architecture of the network was modified to predict speakers from utterances of variable lengths. This was achieved by including a time-delay architecture to operate at the frame level and a statistics pooling layer which collected the

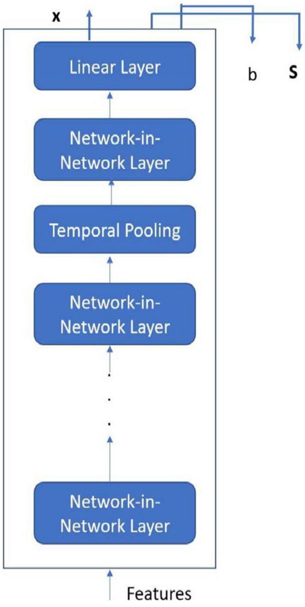


**Fig. 12** X-vector: End-to-end text-independent speaker verification network for enrollment and verification (Snyder et al. 2016)

aggregate activations of the frame-level output layer to pass on to the segment-level part of the network architecture. The output layer of the segment-level network provided the speaker embeddings. Based on the activations of the speaker embeddings, the final softmax layer computed the probability of the kth speaker given a set of $T$ input frames. To enable speaker verification with utterances having variable lengths, the objective function of the optimization problem is computed as,

$$E = -\sum_{n=1}^{N} \sum_{k=1}^{K} d_{nk} \ln\{P(\text{speaker}_k \mid x_{1:T}^{(n)})\}. \tag{8}$$

If there are $K$ speakers in $N$ training segments, then the factor $d_{nk} = 1$ if the $k^{\text{th}}$ speaker gets identified in the nth segment. The network also gives the flexibility to get speaker embeddings from any layer in the segment-level network. Multiple embeddings, which are obtained from the DNN layers, can be combined by concatenation and subjected to their own PLDA backend. The extracted embeddings are known as x-vectors. During experimentation with the NIST SRE 2010 dataset, it was found that in the case of shorter testing utterances, the DNN speaker embeddings were more effective than i-vectors. In the case of the NIST SRE 2016 dataset, the embeddings showed lower EER when testing trails were Cantonese or Tagalog for the English enrollment utterances.

In 2020, Desplanques et al. further refined the x-vector framework to improve the speaker verification performance (Desplanques et al. 2020). The proposed framework was known as emphasized channel attention, propagation, and aggregation (ECAPA) TDNN. With respect to the architecture of the x-vector and ResNet system, three major concerns were identified and resolved by refining the x-vector system architecture. The proposed refinements are listed below:

- The channel and temporal aggregation strategy was modified to attention pooling from the statistical pooling. This enabled the framework to learn speaker information-specific weights for different time frames and channels.
- Instead of normal TDNN layers, the ResNet squeeze excitation ($SE - Res2Block$) blocks were introduced to capture channel inter-dependency and global context information.
- An aggregation strategy was introduced to combine various hierarchical information captured from different layers of the network

The experiments were carried out by considering the Vox Celeb 2 dataset for training and Vox SRC 2019 for testing. The study showed the importance of each modified component. The ECAPA-TDNN framework provided an average improvement of 19% in terms of EER over the existing baseline systems that used Vox SRC 2019 for testing.

Apart from the above, inspired by other speech-based applications, deep learning frameworks that use generative and unsupervised approaches have also been explored for speaker recognition in the recent past. In (Chen et al. 2020), for example, a speaker-conditional generative adversarial network (GAN) framework was used to perform the speaker identification task. It was observed that due to the nature of the learning strategy, generatively capturing the speaker's characteristics through the generator and discriminating them through the discriminator helps the framework to improve performance with less training data. Further, it was observed that the framework performs well in short testing utterances of 1–2 s. In another work, Jati et al. proposed an unsupervised framework called neural predictive coding to perform the speaker verification task (Jati and Georgiou 2019). The framework used the assumption of short-term speaker stationary to extract embeddings and used the Seamase network with contrastive learning to learn speaker discrimination. For contrastive learning, the true pairs are generated by extracting embeddings from the consecutive 1 s segments of the same utterance, and the false pairs segments were sampled from different utterances. The performance of the system was better than the i-vector and x-vector, when supervised in-domain training data was not available. In (Ravanelli et al. 2018), Ravanelli and Bengio proposed a signal-processing-inspired CNN-based framework called Sinc Net to perform the speaker identification and verification task. The CNN filter coefficients were replaced with the sinc function co-coefficients, with only two hyperparameters of minimum and maximum frequency. The sinc filters were hypothesized as sub-band filters with a lower and higher cutoff frequency and were expected to learn the task-specific cutoff frequency during training. The advantage of this approach is it can work well with only 12–16 s of speaker-specific training data and 2–6 s of testing data. The best thing about this framework was its interpretability and the working ability on raw speech signals. This work showed that trained sinc filters capture the speaker-specific formant and pitch information, which can be used for speaker discrimination during testing. However, in realistic scenarios, till today the ECAPA-TDNN-based system remains the state-of-the-art speaker verification framework Fig. 12.

Based on the review of recent developments in the speaker verification systems since the last decade, the following events may be highlighted:

- Introduction of JFA helped to factorize the utterance-dependent test supervectors to its channel/session matrices to get vectors which can be used to discriminate between the speaker classes.
- I-vector with PLDA backend was the defacto speaker verification system till the mid-2010s, for speaker representation in session and channel variabilities. In the availability of longer testing trails, i-vector-based speaker verification systems are still found to exhibit the lowest EER.
- X-vectors derived from ECAPA-TDNN are found to be efficient and the state-of-the-art speaker verification framework in terms of EER.

As is evident from our discussion till now, the NIST SRE corpuses have became the basis for proper comparison between different systems. Each NIST SRE corpus, however, is tailored for a particular challenge, and is not a simple expansion of the previous NIST SRE corpus. Hence, the comparison of different systems on a single large NIST SRE corpus may not be available. Nevertheless, for the benefit of the readers, in Fig. 13, a comparison of a few popular speaker recognition systems has been showcased, courtesy of Brno University of Technology (Matějka et al. 2020). Interested readers are advised to refer to (Matějka et al. 2020) for the specific details of the datasets, evaluation criteria and implementation procedures of the speaker recognition systems.

### 10.2 Recent trends of exploratory research in speaker verification

In the recent past, along with efforts to develop robust systems, a shift towards the exploration of relatively new areas which use/exploits the speaker identity has been observed. Adversarial attack on the speaker verification system deployed for authentication is one such problem. There is also a growing concern for the privacy preservation of the speaker's identity in speech utterances. With reference to certain applications, such as

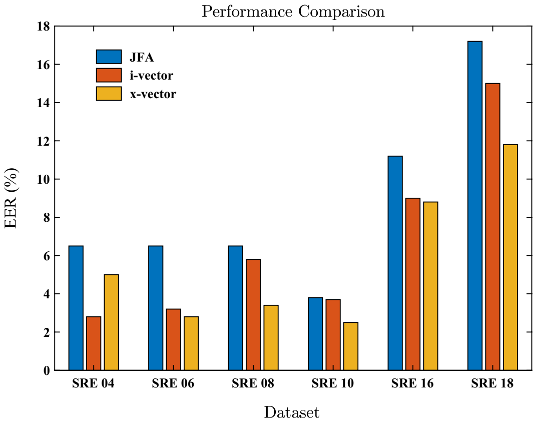


**Fig. 13** Performance comparison of popular speaker recognition systems on different NIST SRE corpuses (Matějka et al. 2020)

speaker identity in medical data (conversation between patient and medical practitioner), anonymizing the speaker identity without affecting other linguistic information is essential (Tomashenko et al. 2022). Therefore, the development of tools for preserving voice privacy has emerged as another thrust area of research related to speaker recognition. The research conducted in these new domains is generally reported in various challenges organized as a part of renowned conferences such as INTERSPEECH, spoken language technology (SLT), and so on. In the following, we summarize some of the important recent challenges relating to these new explorations.

### 10.2.1 Speaker verification in adverse scenarios

Towards including more variabilities, NIST SRE 2016 released the speakers in the wild (SITW) database for speaker verification (McLaren et al. 2016). The SITW database used web-crawled speech data from open-source media (as partial excerpts from the audio recordings), natural speech-degrading artefacts such as noise and other compression formats. The SITW speech data included 299 well-known public figures, or persons of interest (POI), with nearly eight sessions of data for each speaker. The database has a considerable mismatch in audio conditions such as multi-speaker audio from both professionally edited interviews (such as quiet-set interviews, red-carpet interviews, and question-and-answer sessions in an auditorium) as well as more casual, conversational multi-speaker audio in which back-channel, laughter, and overlapping speech were present. Therefore, the metadata of the database has been transcribed with gender, sensor (Mic type), number of speakers, observed artefacts (noise, reverberation, compression), level of degradation for the most prominent artefact, and recording environment. For the SRE SITW challenge, the dataset has been released with two enrollment conditions, namely core enrollment (single speaker) and assist enrollment (multi-speaker enrollment). Individual speaker utterances having a duration from 60-180 s are used in the former case, and annotated conversation speech ranging from 15 s to 1 h is used in the latter. Even though speaker data for testing had been taken from environments which were similar to that of enrollment, the degree of variabilities present is much higher.

Three DNN-based systems and one GMM-UBM i-vector system obtained noteworthy attention in the SITW NIST SRE 2016 challenge. The best system, proposed by Lei et al., replaced the GMM-UBM i-vector system with the DNN i-vector system where i-vectors were defined for each of the phonemic classes (Lei et al. 2016). The second-best system, by Matejka et al., used the standard bottle-neck features (for large vocabulary speech recognition) for speaker verification (Matejka et al. 2014). The third-best system also used DNN-based speaker modelling, replacing the conventional UBM systems for deriving the i-vectors (McLaren et al. 2015). The speaker verification system developed by McLaren et al., in 2011, also performed competitively in the challenge (McLaren and van Leeuwen 2011). In this system, the i-vectors derived from the UBM were projected on the normalized LDA space generated during the speaker enrollment stage, which reduced intra-speaker variabilities.

Based on the NIST SRE evaluation carried out on the SITW database, the following inferences may be made:

- The speech recording of well-known speakers collected from internet sources will have overlapping phonetic content (due to their profession).
- Systems leveraging phonetic features showed better performance compared to the traditional speaker recognition systems.

Since the SITW NIST SRE 2016 challenge, the domain of research has only grown. Interested readers are advised to look separately into this domain (Khosravani and Hoayounpour 2017; Mishra et al. 2022; Ferrer et al. 2022).

### 10.2.2 Speaker verification against spoofing attacks

Speaker verification systems are, in general, robust to zero-effort imposters. However, it is well acknowledged that these systems can be fooled or spoofed, which severely degrades the system performance (Ratha et al. 2001; De Leon et al. 2012). In the literature, the spoofing attacks are broadly categorised as impersonation, replay, text-to-speech (TTS) synthesis and voice conversion (VC) (Wu et al. 2015). Impersonation is performed by a professional who is aware of the target speaker's speech and could mimic the prosody, accent, lexicon, pronunciation, and other important speech properties (Hautamäki et al. 2013; Farrús et al. 2008). The replay attack is carried out by using the pre-recorded speech of the target speaker (Villalba et al. 2011). The TTS synthesis-based spoofing is carried out using the machine-generated speech of the target speaker. Due to recent advancements in TTS methods such as DNN-based TTS and Generative Adversarial Networks (GANs) (Saito et al. 2017), Tacotron (Wang et al. 2017), Wavenet (Oord et al. 2016), etc., a high-quality speech similar to natural speech containing prosodic content can be generated. In VC-based spoofing, a person's speech is converted to speech similar to the target speaker's (Kinnunen et al. 2012). Various sophisticated methods for VC have been developed in recent years which has made this possible (Zhang et al. 2015; Desai et al. 2009; Oord et al. 2016; Saito et al. 2017).

Initially, for anti-spoofing research, different speech and speaker recognition databases, such as YOHO (Lau et al. 2004), NIST (Hautamäki et al. 2013), and WSJ (De Leon et al. 2012) were used. In the special session of INTERSPEECH 2013, the need for a common dataset, protocol and metrics was for the first time realised and hence a speaker verification spoof challenge was organised in INTERSPEECH 2015 (Wu et al. 2015). The database provided in the challenge contains genuine and spoofed speech. The genuine speech consists of 106 human speaker's speech (45 male and 61 female) having no significant channel or background noise effects. Spoofed speech is the modified form of genuine speech using the TTS and VC methods. The details of the generation of spoofed speech are available in (Wu et al. 2015). Another spoofing database was provided in the 'Biometrics: Theory, Applications, and Systems' (BTAS) 2016 challenge, which contains replay spoofing attacks along with TSS and VC spoofing attacks (Ergünay et al. 2015). The BTAS 2016 dataset has 'unknown' attacks in the test set to make it comparatively more challenging. One more spoofing database, focused on replay spoofing, was developed from the re-recording of the original RedDots database. The database was developed from replay attacks having recording and playback conditions. The RedDots database consists of genuine speech and its replay version as spoofed data. The second edition of the speaker verification spoof challenge was conducted in Interspeech 2017 (Kinnunen et al. 2017). The challenge focused on the detection of only replay attacks which can be done very easily with the help of high-quality recording devices.

Similar to the case of the SITW NIST SRE 2016, the BTAS 2016 challenge led to the growth and visibility of this domain of research. Interested readers are advised to look separately into this domain (Kinnunen et al. 2017; Nautsch 2021; Yamagishi et al. 2021; Jung et al. 2022).

## 11 Summary

Research in speaker recognition began in the 1960s, almost a decade after the onset of speech recognition technology. During the early days, research was limited in scope and capacity, which inevitably expanded with the advancement of computing technology. Consequently, it is easy to summarize the key events of the 1960s and 70s, and, in comparison, quite challenging to summarize the recent past. Initially, researchers explored all the different possibilities in speaker recognition: speaker identification, speaker verification, text-dependent and text-independent. As time progressed, however, as discussed in Section 7, text-independent speaker verification came to be recognised as the most promising sub-domain of speaker recognition. The 1990s and 2000s led to many key technological advancements, such as the GMM, JFA and score-normalization techniques, and they have been given adequate and separate attention in this article. Finally, the last and the present decade, which has developed on the back of the hard work done in the 1990s and 2000s, have been summarized, highlighting the key breakthroughs, such as the i-vector and x-vector sytems, and other AI based systems.

It is only since the onset of the new millenium that there has been significant and widespread international collaboration and competition in academic research. As such, only since the 2000s, researchers have been able to implement and evaluate different speaker recognition systems on significantly large datasets. As discussed in Section 7, in 2002, Reynolds compared four speaker verification systems: but the databases used for each of the systems were different, as they catered to four different problems. In the recent past, different NIST SRE corpuses, tailored for different challenges, have become the basis for performance evaluation. For the benefit of the readers, in Fig. 13, at the end of Section 10.1, a comparison of a few popular speaker recognition systems has been showcased.

## 12 Conclusion

This article provides a historical overview of research in the domain of speaker recognition, focusing on the key works of every decade. While many exhaustive review articles exist, the objective of this article is to be selective and less overwhelming, highlighting works which represent the pattern of thinking in the research community of that particular time. The nature of this document requires a lot of significant work to be left out, inevitably. What is significant and what is not is also a subjective question, and hence this document is by no means perfect. However, as a whole, this document should serve as a concise yet comprehensive guide to the engineers and researchers of the present and the future.

**Funding** This research was conducted as part of the sub-project, "Development and Deployment of Scalable Speaker Recognition Technology in Different Applications," inclusive of the main-project, "Speech Technologies in Indian languages". The projects are supported and funded under the National Language Translation Mission (NLTM) of the Ministry of Electronics and Information Technology (MeitY), Government of India.

**Code availability** Not applicable.

## Declarations

**Conflict of interest** The authors have no conflicts of interest to declare that are relevant to the content of this article.

## References

Aharon M, Elad M, Bruckstein A (2006) K-svd: an algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans Signal Process 54(11):4311–4322

Atal B (1972) Text-independent speaker recognition. J Acoust Soc Am 52(1A):181–181

Atal BS (1974) Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. J Acoust Soc Am 55(6):1304–1312

Bai Z, Zhang X-L (2021) Speaker recognition based on deep learning: an overview. Neural Netw 140:65–99

Beek B, Neuberg E, Hodge D (1977) An assessment of the technology of automatic speech recognition for military applications. IEEE Trans Acoust, Speech, Signal Process 25(4):310–322

Benesty J, Sondhi MM, Huang Y (2008) Springer handbook of speech processing. Springer, Berlin

Bimbot F, Bonastre J-F, Fredouille C, Gravier G, Magrin-Chagnolleau I, Meignier S, Merlin T, Ortega-García J, Petrovska-Delacrétaz D, Reynolds DA (2004) A tutorial on text-independent speaker verification. EURASIP J Adv Signal Process 2004(4):1–22

Bricker P, Gnanadesikan R, Mathews M, Pruzansky S, Tukey P, Wachter K, Warner J (1971) Statistical techniques for talker identification. Bell Syst Tech J 50(4):1427–1454

Brummer N, Burget L, Cernocky J, Glembek O, Grezl F, Karafiat M, Van Leeuwen DA, Matejka P, Schwarz P, Strasheim A (2007) Fusion of heterogeneous speaker recognition systems in the stbu submission for the nist speaker recognition evaluation 2006. IEEE Trans Audio, Speech, Lang Process 15(7):2072–2084

Campbell WM, Campbell JP, Reynolds DA, Singer E, Torres-Carrasquillo PA (2006) Support vector machines for speaker and language recognition. Comput Speech Lang 20(2–3):210–229

Campbell WM, Sturim DE, Reynolds DA (2006) Support vector machines using gmm supervectors for speaker verification. IEEE Signal Process Lett 13(5):308–311

Chen L, Liu Y, Xiao W, Wang Y, Xie H (2020) Speakergan: speaker identification with conditional generative adversarial network. Neurocomputing 418:211–220

Chen K, Salman A (2011) Learning speaker-specific characteristicswith a deep neural architecture. IEEE Trans Audio, Speech Lang Process 22(11):1744–1756

Davis S, Mermelstein P (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoust, Speech, and Signal Processing 28(4):357–366

Dehak N, Kenny P, Dehak R, Dumouchel P, Ouellet P (2010) Front-end factor analysis for speaker verification. IEEE Trans Audio Speech Lang Process 19(4):788–798

Desai S, Raghavendra EV, Yegnanarayana B, Black AW, Prahallad K (2009) Voice conversion using artificial neural networks, In: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2009, pp. 3893–3896

Desplanques B, Thienpondt J, Demuynck K (2020) ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification, In: Proc. Interspeech 2020, pp. 3830–3834

Dey S, Barman S, Bhukya RK, Das RK, Haris BC, Prasanna SRM, Sinha R (2014) Speech biometric based attendance system, In National Conference on Communications

Doddington GR (1971) A method or speaker verification. J Acoust Soc Am 49(1A):139–139

Doddington G (2001) Speaker recognition based on idiolectal differences between speakers, In: Seventh European Conference on Speech Communication and Technology

Endres W, Bambach W, Flösser G (1971) Voice spectrograms as a function of age, voice disguise, and voice imitation. J Acoust Soc Am 49(6B):1842–1848

Ergünay S. K, Khoury E, Lazaridis A, Marcel S (2015) On the vulnerability of speaker verification to realistic voice spoofing, in 2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS). IEEE, 2015, pp. 1–6

Farrús M, Wagner M, Anguita J, Hernando J(2008) How vulnerable are prosodic features to professional imitators? in Odyssey 2008: The Speaker and Language Recognition Workshop; 2008 Jan. 21-24; Stellenbosch (South Africa).[place unknown]: ISCA; 2008. Paper 002 [4 p.]. International Speech Communication Association (ISCA),

Faundez-Zanuy M, Monte-Moreno E (2005) State-of-the-art in speaker recognition. IEEE Aerosp Electron Syst Mag 20(5):7–12

Ferrer L, McLaren M, Brummer N (2022) A speaker verification backend with robust performance across conditions, computer. Speech Lang 71:101258

Furui S (1981) Cepstral analysis technique for automatic speaker verification. IEEE Trans Acoust, Speech, Signal Process 29(2):254–272

Furui S (2005) 50 years of progress in speech and speaker recognition research. ECTI Trans Comput Inform Technol 1(2):64–74

George KK, Kumar CS, Sivadas S, Ramachandran KI, Panda A (2018) Analysis of cosine distance features for speaker verification. Pattern Recognit Lett 112:285–289

Gupta POV, Kenny PST (2014)I-vectorbased speaker adaptation of deep neural networks forfrench broadcast audio transcription, In: Proc. ICASSP

Hanifa RM, Isa K, Mohamad S (2021) A review on speaker recognition: technology and challenges. Comput Electr Eng 90:107005

Haris BC, Pradhan G, Sinha R, Prasanna SRM (2013) The iitg speaker verification systems for nist sre 2012, In Proc. ICASSP

Haris BC (2014) Joint sparse coding over learnt dictionaries and the use of low complexity projections for speaker verification, Ph.D. dissertation, Indian Institute of Technology Guwahati

Haris BC, Sinha R (2011) Exploring sparse representation classificationfor speaker verification in realistic environment, In: Proc. Cenenary conference in Electrical Engineering

Haris BC, Sinha R (2012) Speaker verification using sparserepresentation over ksvd learned dictionary, In: Proc. National Conference on Communication

Hautamäki RG, Kinnunen T, Hautamäki V, Leino T, Laukkanen A-M (2013) I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry. In Interspeech. Citeseer: 930–934

Heigold G, Moreno I, Bengio S, Shazeer N(2016) End-to-end text-dependent speaker verification, In: Proc. ICASSP

Higgins A, Bahler L, Porter J (1991) Speaker verification using randomized phrase prompting. Digit Signal Process 1(2):89–106

Jati A, Georgiou P (2019) Neural predictive coding using convolutional neural networks toward unsupervised learning of speaker characteristics. IEEE/ACM Trans on Audio, Speech, Lang Process 27(10):1577–1589

Jung J-W, Tak H, Jin Shym H, et al. (2022) Sasv 2022: The first spoofing-aware speaker verification challenge, In: Proc. Interspeech

Kabir MM, Mridha MF, Shin J, Jahan I, Ohi AQ (2021) A survey of speaker recognition: fundamental theories, recognition methods and opportunities. IEEE Access 9:79–236

Kenny P, Ouellet P, Dehak N, Gupta V, Dumouchel P (2008) A study of interspeaker variability in speaker verification. IEEE Trans Audio, Speech, Lang Process 16(5):980–988

Kenny P (2005) Joint factor analysis of speaker and session variability: Theory and algorithms, CRIM, Montreal,(Report) CRIM-06/08-13, 14(28-29):2,

Kenny P (2010) Bayesian speaker verification with heavy-tailedpriors, In Proc, Odyssey

Kenny P, Gupta V, Stafylakis T, Ouellet P, Alam J (2014) Deepneural networks for extracting baum-welch statistics for speakerrecognition, In Proc. Odeyssey

Khosravani A, Hoayounpour MM (2017) A plda approach for language and text independent speaker recognition. Comput Speech Lang 45:457–474

Kinnunen T, Li H (2010) An overview of text-independent speaker recognition: from features to supervectors. Speech Commun 52(1):12–40

Kinnunen T, Sahidullah M, Delgado H, Todisco M, Evans N, Yamagishi J, Lee K. A (2017) The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection, In: Proc. Interspeech 2017, pp. 2–6

Kinnunen T, Lee K-A, Li H (2008) Dimension reduction of the modulation spectrogram for speaker verification. In Odyssey, p. 30

Kinnunen T, Wu Z.-Z, Lee K. A, Sedlak F, Chng E. S, Li H (2012) Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech, in 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4401–4404

Krishnamoorthy P, Jayanna HS, Prasanna SRM (2011) speaker recognition under limited data condition by noise addition. Exp Syst Appl 38(10):13487–13490

Lau YW, Wagner M, Tran D (2004) Vulnerability of speaker verification to voice mimicking, In: Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004. IEEE, 2004, pp. 145–148

Lei Y, Scheffer N, Ferrer L, McLaren M (2016) A novel scheme for speaker recognition using a phonetically-aware deep neural network In Proc. INTERSPEECH

De Leon PL, Pucher M, Yamagishi J, Hernaez I, Saratxaga I (2012) Evaluation of speaker verification security and detection of hmm-based synthetic speech. IEEE Trans Audio, Speech, Lang Process 20(8):2280–2290

Li K-P, Hughes G (1974) Talker differences as they appear in correlation matrices of continuous speech spectra. J Acoust Soc Am 55(4):833–837

Lu X, Dang J (2008) An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification. Speech Commun 50(4):312–322

Ma B, Zhu D, Tong R, Li H (2006) Speaker cluster based gmm tokenization for speaker recognition. In Interspeech

Markel J, Oshika B, Gray A (1977) Long-term feature averaging for speaker recognition. IEEE Trans Acoust, Speech, and Signal Process 25(4):330–337

Matejka P, Zhang L, Ng T, Glembek O, Ma JZ, Zhang B, Mallidi SH (2014) Neural network bottleneck features for language identification, In: Proc. Speaker and language recognition workshop, Odyssey

Matsui T, Furui S (1994) Comparison of text-independent speaker recognition methods using vq-distortion and discrete/continuous hmm's. IEEE Trans Speech Audio Process 2(3):456–459

Matsui T, Furui S (1993) Concatenated phoneme models for text-variable speaker recognition, In: 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE, 2:391–394

Matsui T, Furui S (1994) Similarity normalization method for speaker verification based on a posteriori probability, In: Proc. ESCA Workshop on Automatic Speaker Recognition Identification Verification. Switzerland, pp 59-62

Matějka P, Plchot O, Glembek O, Burget L, Rohdin J, Zeinali H, Mošner L, Silnova A, Novotnỳ O, Diez M et al (2020) 13 years of speaker recognition research at but, with longitudinal analysis of nist sre. Comput Speech Lang 63:101035

McLaren M, Ferrer L, Castan D, Lawson A (2016) The speakers in thewild (sitw) speaker recognition database, In: Proc. INTERSPEECH

McLaren M, Lei Y, Ferrer L(2015) Advances in deep neural network approaches to speaker recognition, In: proc. icassp 2015, In: Proc. ICASSP

McLaren M, van Leeuwen D (2011) Source -normalised-and-weighted lda for robust speaker recognition using i-vectors, In: Proc. ICASSP

Mishra J, Bhattacharjee M, Prasanna S. R. M (2022) Imsv 2022: indic multi lingual and multi sensor speaker verification challenge, In: Proc. O-COCOSDA

Naik JM, Netsch LP, Doddington GR (1989) Speaker verification over long distance telephone lines, In: International Conference on Acoustics, Speech, and Signal Processing, IEEE, 1989:524–527

Nakagawa S, Wang L, Ohtsuka S (2011) Speaker identification and verification by combining mfcc and phase information. IEEE Trans Audio, Speech, Lang Process 20(4):1085–1095

Nautsch A et al (2021) Asv spoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech. IEEE Trans Biom, Behav Identity Sci 3(2):252–265

Novotny O, Matejka P, Glembeck O, Plchot O, Grèzl F, Burget L, Cernocky J (2016) Analysis of the dnn-based sre systemsin multi-language conditions, in in Proc. Spoken Language Technology Workshop (SLT)

Oord Avd, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A, Kavukcuoglu K (2016) Wavenet: a generative model for raw audio, arXiv preprint arXiv:1609.03499,

Pati Y, Rezaiifar R, Krishnaprasad P (1993) Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition, In: Proc. 27th Asilomar Conference on Signals, Systems and Computers

Poritz A (1982) Linear predictive hidden markov models and the speech signal, in ICASSP'82. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, IEEE 1982:1291–1294

Pradhan G, Prasanna SRM (2013) Speaker verification by vowel and nonvowel like segmentation. IEEE Trans Audio, Speech Lang Process 21(4):854–867

Prasanna SM, Gupta CS, Yegnanarayana B (2006) Extraction of speaker-specific excitation information from linear prediction residual of speech. Speech Commun 48(10):1243–1261

Prasanna SRM, Pradhan G (2011) Significance of vowel-like regions for speaker verification under degraded conditions. IEEE Transon Audio, Speech Lang Process 19(8):2552–2565

Pruzansky S, Mathews MV (1964) Talker-recognition procedure based on analysis of variance. J Acoust Soc Am 36(11):2041–2047

Ratha NK, Connell JH, Bolle RM (2001) Enhancing security and privacy in biometrics-based authentication systems. IBM Syst J 40(3):614–634

Ravanelli M, Bengio Y (2018) Speaker recognition from raw waveform with sincnet, In: 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE 2018:1021–1028

Reynolds DA (1995) Speaker identification and verification using gaussian mixture speaker models. Speech Comm 17(1–2):91–108

Reynolds DA, Quatieri TF, Dunn RB (2000) Speaker verification using adapted gaussian mixture models. Digit Signal Process 10(1–3):19–41

Reynolds DA (2002) An overview of automatic speaker recognition technology, In: 2002 IEEE international conference on acoustics, speech, and signal processing 4: IV–4072

Rose RC, Reynolds DA (1990) Text independent speaker identification using automatic acoustic segmentation, In: International Conference on Acoustics, Speech, and Signal Processing. IEEE, 1990: 293–296

Rosenberg A, Sambur M (1975) New techniques for automatic speaker verification. IEEE Trans Acoust, Speech, Signal Process 23(2):169–176

Rosenberg AE, Soong FK (1987) Evaluation of a vector quantization talker recognition system in text independent and text dependent modes. Comput Speech Lang 2(3–4):143–157

Saito Y, Takamichi S, Saruwatari H (2017) Statistical parametric speech synthesis incorporating generative adversarial networks. IEEE/ACM Trans Audio, Speech, Lang Process 26(1):84–96

Sambur MR (1973) Speaker recognition and verification using linear prediction analysis. J Acoust Soc Am 53(1):354–354

Senoussaoui M, Kenny P, Pierre Dumouchel ND (2010) An i-vector extractor suitable for speaker recognition with both microphone and telephone speech, In: The speaker and Langugae Recognition Workshop

Sharma R, Bhukya RK, Prasanna SM (2018) Analysis of the hilbert spectrum for text-dependent speaker verification. Speech Commun 96:207–224

Sharma R, Prasanna SM, Bhukya RK, Das RK (2017) Analysis of the intrinsic mode functions for speaker information. Speech Commun 91:1–16

Shriberg E, Ferrer L, Kajarekar S, Venkataraman A, Stolcke A (2005) Modeling prosodic feature sequences for speaker recognition. Speech Commun 46(3–4):455–472

Snyder D, Ghahremani P, Povey D, Garcia-Romero D, Carmiel Y, Khudanpur S (2016) Deep neural network-based speaker embeddings for end-to-end speaker verification, in in Proc. Spoken Language Technology Workshop

Snyder D, Garcia-Romero D, Povey D, Khudanpur S (2017) Deep neural network embeddings for text-independent speaker verification, In Proc. INTERSPEECH

Soong FK, Rosenberg A, Rabiner L, Juang B (1987) A vector quantization approach to speaker recognition. ATT Tech J 66:22

Stafylakis T, Kenny P, Senoussaoui M, Dumouchel P(2012) Preliminary investigation of boltzmann machine classifiersfor speaker recognition, In Odyssey

Strang G (2009) Introduction to Linear Algebra. Wellesley-Cambridge Press, Cambridge

Tisby NZ (1991) On the application of mixture ar hidden markov models to text independent speaker recognition. IEEE Trans Signal Process 39(3):563–570

Tomashenko N, Wang X, Vincent E, Patino J, Srivastava BL, Noe P-G, Nautsch A, NicholasEvans Yamagishi J, O'Brien B, Chanclu A, Bonastrea J-F, Todiscod M, Maouch M (2022) The voiceprivacy 2020 challenge: results and findings. Comput Speech Lang 74(10132):1–38

Variani E, Lei X, McDermott E, Moreno I. L, Gonzalez-Dominguez J(2014) Deep neural networks fr small foot print text-dependent speaker verification, In: Proc. ICASSP

Vasilakakis V, Cumani S, Laface P (2013) Speaker recognition by means of deep belief networks, in in Proc. Biometric Technologiesin Forensic Science,

Villalba J, Lleida E, Detecting replay attacks from far-field recordings on speaker verification systems, in Biometrics and ID Management: COST 2101 European Workshop, BioID, (2011) Brandenburg (Havel), Germany, March 8–10, 2011. Proceedings 3. Springer 2011:274–285

Wang Y, Skerry-Ryan R, Stanton D, Wu Y, Weiss R. J, Jaitly N, Yang Z, Xiao Y, Chen Z, Bengio S, et al.,(2017) Tacotron: Towards end-to-end speech synthesis, arXiv preprint arXiv:1703.10135,

Wu Z, Evans N, Kinnunen T, Yamagishi J, Alegre F, Li H (2015) Spoofing and countermeasures for speaker verification: a survey. Speech Commun 66:130–153

Wu Z, Kinnunen T, Evans N, Yamagishi J, Hanilçi C, Sahidullah M, Sizov A (2015) ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge, In: Proc. Interspeech 2015, , pp. 2037–2041

Wu Z, Khodabakhsh A, Demiroglu C, Yamagishi J, Saito D, Toda T, King S, Sas: A speaker verification spoofing database containing diverse attacks, In: (2015) IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE 2015:4440–4444

Yamagishi J, Wang X, Todisco M, et al. (2021) Asv spoof 2021: Accelerating progress in spoofed and deepfake detection, In: Proc. Interspeech

Zhang S, Huang D, Xie L, Chng ES, Li H, Dong M (2015) Non-negative matrix factorization using stable alternating direction method of multipliers for source separation, In: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). IEEE 2015:222–228

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.