Peertechz

ENGINEERING GROUP

TRENDS IN
Computer Science and Information Technology

ISSN: 2641-3086 DOI: https://dx.doi.org/10.17352/tcsit

Junxia Liu[1], CL Philip Chen[1,2]*, Tieshan Li[1]*, Yi Zuo[1] and Peichao He[1]

[1]Dalian Maritime University, Dalian 116026, China
[2]University of Macau, Macau 99999, China

Received: 19 July, 2019
Accepted: 26 August, 2019
Published: 28 August, 2019

***Corresponding authors:** CL Philip Chen and Tieshan Li, Department of Navigation Institute, DaLian Maritime University, Linghai Road Ganjingzi District, Dalian, Liaoning, Room 518, China, Tel: 010-0411-84729255;
E-mail: philip.chen@ieee.org; tieshanli@126.com

Keywords: Speaker recognition; Feature extraction; MFCC; Deep learning; End-to-end model

https://www.peertechz.com

Review Article

An overview of speaker recognition

Abstract

Speaker recognition has been studied for many years and has been a hot topic. This paper presents an overview of speaker recognition methods, which include the classical and the state-of-art methods. According to the modular components of speaker recognition system, we firstly introduced the fundamentals of speaker recognition, which are mainly divided into two parts: feature extraction and speaker modeling. The most commonly speech features used in speaker recognition were elaborated firstly. In particular, the recent progress of deep neural network proposes a new approach of feature extraction and has become the technology trend. Secondly, the classical approaches of speaker recognition model were introduced, and elaborated the recent progress of deep learning speaker recognition. This paper especially provides an in-depth analysis on end-to-end model which consists of a training component to extract features, an enrollment component to training the speaker model, and an evaluation component with appropriate loss function for optimization. The final part concludes the paper with discussion on future trends.

Introduction

Speaker recognition bas become the most popular methods in biometric identification field, because the voice is the most common signal and the simplest to acquire [1,2]. With the wide application of artificial intelligence machines, researchers have found that voice communication is the best way to communicate between humans and machines. In access control field, telephone services of transaction authorization field, and speaker diarization field, speaker recognition has been applied extensively [3,4]. In general, the speaker recognition system can fall into two categories: speaker identification (SI) and speaker verification (SV). Speaker identification is the process to determining who is talking from a group of people, and the system must perform a 1: N classification. Speaker verification is the task to determining whether a person is who he/she claims to be (a yes/no decision). Generally, speaker identification can be divided into "closed-set" and "open-set", it is supposed that the training voice come from a fixed set of know speakers, thus the task is referred to as closed-set identification. Instead, it is assumed that training voice are not known to the system, that is refered to as open-set identification [5-7].

The speech used for speaker recognition can be grouped into text-dependent (TD) and text-independent (TI). In the text-dependent (TD) application, the recognition system has prior knowledge of the text to be spoken and it is expected the recognition must be pronounced according to the prior knowledge. Because of prior knowledge, the text-dependent (TD) recognition can greatly improve performance of recognition system. In a text-independent (TI) application, there is no fixed text for the recognition system to be spoken. Since there is no prior knowledge, text-independent speaker recognition is more difficult but also more flexible. As the speech recognition accuracy improved and the speaker and speech recognition system rapid develop, the distinction between text-independent and text-dependent application have been decreased [8-13].

The methods of speaker recognition's research and development have been studied over five decades, which still is an active area. The methods of speaker recognition are spanned from the original human aural spectrogram comparisons to simple template matching, to dynamic time-warping approaches, to more modern statistical pattern recognition approach and to the most popular deep learning in recent years. [14-18]. Especially noting that, the methods applied to speech recognition have also been often used in speaker recognition (SR). The corpora' research and development are from small, private corpora to large, open source corpora. This domain has natured to the degree that commericial applications of SR have been growing steadily since the mid-1980s, and many large companies have this technology, such as Google, Baidu, IBM and Microsoft etc have set up speech research groups.

This paper is a general overview of speaker recognition technologies, that introduce the classic techniques from 1987 until today. Meanwhile, we focus on the recent techniques that shifted from deep neural network models to end-to-end models. The remaining of this overvies is organized as follow: section

Citation: Liu J, Philip Chen CL, Li T, Zuo Y, He P (2019) An overview of speaker recognition. Trends Comput Sci Inf Technol 4(1): 001-012.
DOI: http://dx.doi.org/10.17352/tcsit.000009

2 introduce the development of speaker recognition. Section 3 introduce fundamentals of speaker recognition. Section 4 and 5 elaborate feature extraction and speaker modeling process. Section 6 is then devoted to the decision method. The end-to-end model was introduced with emphasis in section 7. Finally, the conclusion and the future research trends of recognition technology are outlined in Section 8.

Overview

Speaker recognition can be put into four stages. The first stage was from 1960s to 1970s, the research focused on feature extraction and template matching technique. In 1962, Kesta at Bell LABS proposed spectrogram method for speaker recognition [19]. In 1969, Luck proposed Cepstrum technology [20]. In 1976, Atal et al. proposed the Linear Predictive Cepstrum Coefficients (LPCCs), which improved the accuracy of speaker recognition [21]. In terms of the model, template matching was mainly adopted in the 1960s. In the 1970s, Dynamic Time Warping (DTW) and Vector Quantization (VQ) technology became the mainstream.

The second stage was from the 1980s to the 1990s, speech statistical models are beginning to be applied to speaker recognition [22-24]. In terms of feature extraction, Davis proposed Mel-Frequency cepstrum parameter (MFCC) for speaker recognition which becomes the mainstream feature in the following years [25]. In terms of models, the classical approach has been devided into two types. The first types are based on vector quantization and dynamic time wrapping, which are referred to as template-based models. The second types are stochastic models which based on Gaussian Mixture Model (GMM) [26] or Hidden Markov Model (HMM) [27,28]. The majority of the state-of-the-art SR systems adopted MFCC as features and Gaussian mixture model (GMM) was used for speaker modeling [29-31]. The Gaussian Mixture Model has been proved extremely successful in TI speaker recognition.

The third stage around 2000, GMM-based speaker recognition methods has been the most commonly used and which proposed by Reynolds, which include the classical Maximum a-Posteriori (MAP) adaptation of universal background model parameters (GMM-UBM) [32] and support vector machine (SVM) classification of GMM supervectors (GMM-SVM) [33].

In the training phase, the MAP adaptation framework provides a way of incorporating prior information by adapting the parameters of GMM from the UBM. The framework is available in dealing with problems posed by sparse training data [34]. The SVM uses a non-linear function to map data on to a higher (possibly infinite) dimensional space and then finds the best hyper-plane separating the two classes in this space [34,35]. Since the SVM is basically a two-class classifier.

However, in terms of data, the high accurate rate only can be achieved under ideal conditions and is appropriate for practical application under matched channel conditions. Instead, the performance can degrade significantly under mismatched conditions. After 2010,

The barrier associated with compensating for these differences have offered an active research focus for the SV field and some of the most advanced channel compensation schemes include joint factor analysis (JFA) [36], i-vectors [37], or nuisance attribute projection (NAP) [38]. Meanwhile, using the fusion information from different sourec of evidence, which can improve system performance.

The fourth stage began in the early 2000s (2010), deep learning promotes the development of speaker recognition. At this stage, the development of speaker recognition technology is drivern by the commercial needs, while deep learning, big data and genetal graphics computing uint (GPU) also promote the development of speaker recognition. The various deep neural networks based were proposed for speaker cognition methods [39]. At the feature extraction of frame-level, researchers apply deep neural networks to extract Bottlenect (BN) features [40], d-vector [41], j-vector [42], and x-vector [43]. At the model-level, the research focuses on various deep neural networks (DNNs) for acoustic feature modeling. The decisions were maked utilize the distance between the target feature vector and the test feature vector. But speakers are often unknown during system training, this makes a big challenging for SR.

Whether the i-vector system, or feature vector extracted from DNN system, it usually consists of three modules, the first is the training module which calculate the representations of speaker, the second is enrollment module which estimate the speaker model, the last is evaluation module which have an appropriate loss function for optimization. A new approach has been proposed in which all the modules can be jointed together. Compared with the present methods, such an end-to-end(E2E) method direct modeling utterances and directly joint estimation, which result in better and more compact models. Moreover, this approach ofter results in properly simplied systems need fewer concepts and heuristics [44].

Fundamentals

A SR system can typically be divided into three parts as shown in figure 1. The front-end is the processing of the raw speech and then obtaining a set of speaker discriminate features which represent the speaker's characteristics (section 4). The back-end is the modeling and decision-making, training a speaker model using the extracted features (section 5) and decision-logic model is used to produce recognition scores by comparing features from different utterances (section 6). As stated above,

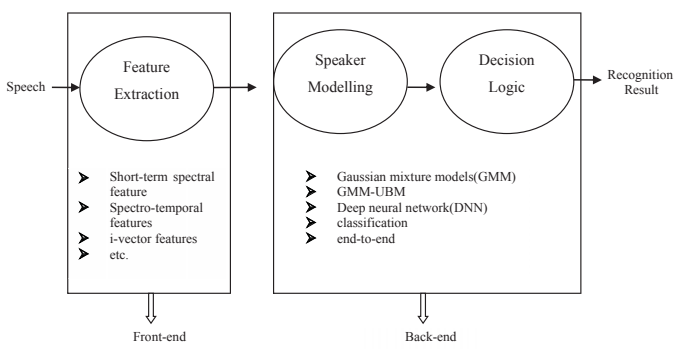

Figure 1: The typical SR system.

Citation: Liu J, Philip Chen CL, Li T, Zuo Y, He P (2019) An overview of speaker recognition. Trends Comput Sci Inf Technol 4(1): 001-012.
DOI: http://dx.doi.org/10.17352/tcsit.000009

the latest end-to-end neural speaker recognition systems were proposed which combining the above two components (front-end and back-end) (section 7).

The basic principles of SR are shown in the figure 2, the top figure is the enrollment process, while the below figure is the recognition process. The function of feature extraction module is to transforming the raw signal into feature vectors. In the enrollment module, the speaker module is trained utilizing the feature vectors of the tagged speaker. In the recognition module, the feature-vectors firstly extracted from the unknown speaker's utterances are compared with the model in the database of system to giving a similarity score. The final decision of SR model is maked using scoring standard.

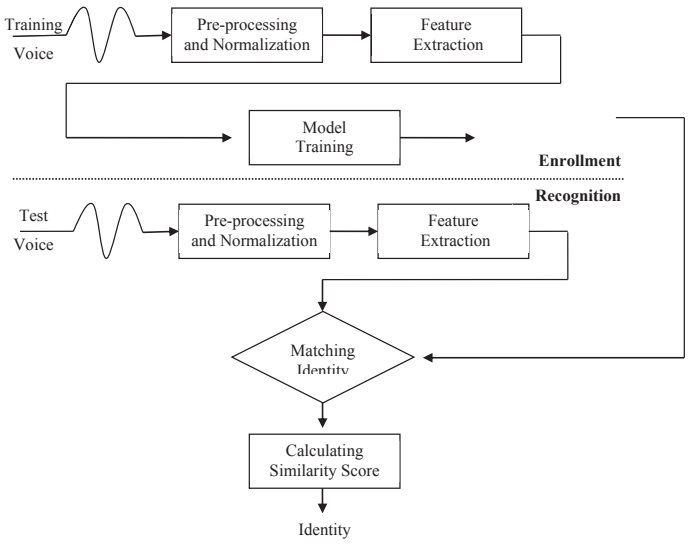

Figure 2: The components of speaker recognition system.

Feature Extraction

The process of feature extraction is to transforming the raw speech signals into some types of abstract expression, namely feature vectors, in which the properties of specific speaker are emphasized. In speaker recognition system, the features can be grouped into two categories: low-level infromation and high-level information. While all of these information conveys useful information for speaker's identity. In the last forty years of speaker recognition, short-term and lower-level acoustic information exclusively is the most useful feature, such as cepstral features. For the high-level information, many researches have investigated the potential benefits of high-level characteristics of speech [45]. In contemporary speaker recognition applications, high-level information needs sufficient training data and very large memory. In the situation of high computational cost, the high-level features received much attention [46]. Hence the most advanced SR system still uses the low-level information. The reporter in this paper focuses on capturing the low-level information by short-term spectral features which are the simplest, yet the most discriminative.

Short-term spectral features

The most commonly features in speaker verification is the Mel frequency cepstral coefficients (MFCCs) [47], Linear prediction cepstral coefficients(LPCCs) [48], and the perceptual linear prediction coefficients(PLPs) [49]. In speaker and speech recognition system, the most fundamental process is that of extracting feature-vector of uniformly spaced across time from the time-domain sampled acoustic waveform, the processes as follows:

1. Pre-emphasis: Pre-emphasis is essentially a high-pass filter which applied to the waveform: $y(t) = x(t) - 0.97x(t-1)$ where $x(t)$ is the input speech data and $y(t)$ is the output. The purpose of pre-emphasis is to emphasises the higher frequencies and flattens the spectrum of the signal. Meanwhile, pre-emphasis can elimanate the effects of vocal cords and lips during vocal production.

2. Framing: The frame is the collection of N sampling points. The purpose of framing is to divided the time-domain waveform into overlapping fixed duration segments, and typical the duration values of a frame is from 20 ms to 30 ms (usually 25 ms). In order to avoid large changes between adjacent frames, there will be an overlap between two adjacent frames. Usually, the values of overlap are about 1/2 or 1/3 of a frame.

3. Windowing: In order to increase the continuity of the left and right sides of the frame, each frame is multiplied by a window function. The window functions usually include hamming window, hanning window, and rectangular window, hamming windows are usually used. Suppose the signal is $S(n), n = 0,1,\cdots N-1, N$ after framing, where the N is the length of a frame. Then each frame is multiplied by hamming windows, $S'(n) = S(n) \times W(n)$, where

$$W(n) = 0.54 - 0.46 \times \cos\left[\frac{2\pi n}{N-1}\right], 0 \le n \le N-1$$

Mel frequcency cepstral coefficients (MFCCs): MFCC has been widely used to capture the speech-specific characteristics for decades in speech processing. MFCC features are derived as follows:

1) The transformation of the signal in the time domain is usually difficult to observe the characteristic of the signal, the frame of N samples in the time domain is transformed to the frequency domain. FFT provides a faster implementation of the Discrete Fourier Transform (DFT). On the setting of N samples DFT will have the following coefficients.

$$X_a(k) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi k/N}, 0 \le k \le N \quad (1)$$

where $x(n)$ is the input speech signal. N is the number of points in the Fourier transform.

2) **Mel frequency warping:** The Mel scale is related to perceived frequency, or pitch of a pure tone to its actual measured frequency. Humans are more sensitive to low-

Citation: Liu J, Philip Chen CL, Li T, Zuo Y, He P (2019) An overview of speaker recognition. Trends Comput Sci Inf Technol 4(1): 001-012.
DOI: http://dx.doi.org/10.17352/tcsit.000009

frequency signals than high frequencies and thus are much better at discerning small changes. Incorporating the Mel-scale makes our features match more closely what humans hear. The formula of Mel-scale is:

$$M(f) = 1125\ln(1 + f/700) \quad (2)$$

The windowed signal spectrum is multiplied with Mel filter bank coefficients.

3) **Discrete Cosine transform:** The MFCC coefficint is obtained by discrete cosine transform(DCT):

$$C(m) = \sum_{m=0}^{N-1} s(m)\cos\left(\frac{\pi m(m-0.5)}{M}\right) \quad n = 1,2,....L \quad (3)$$

The above logarithm energy is introduced into the discrete cosine transform to obtain the Mel-scale Cepstrum parameter of L order, L order is the MFCC coefficient order, L usually take 12-16. Here M is the number of triangular filter. An overview of MFCC computation is shown in figure 3.

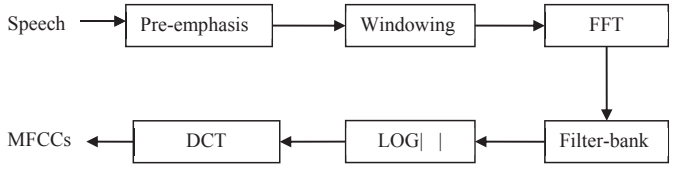

Figure 3: Overview of MFCCfeature extraction.

Linear prediction cepstral coefficients (LPCCs): The linear prediction coefficients attempts to describe a speech signal $s[n]$ at time n as a linear combination of P past signal values as follows

$$\tilde{s}[n] = \sum_{k=1}^{P} a_k s[n-k] \quad (4)$$

where a_k are the linear prediction coefficients. The coefficients determined by minimising the mean-squared prediction error between the speech sample, s[n], and its linearly predicted value, $\tilde{s}[n]$ is determined by the Levinson-Durbin algorithm. While the coefficients of LP model form the basic feature set of SR systems, in order to more sui for speaker modelling and classification, they are usually converted to a more appropriate representation (i.e. cepstral coefficients). LPCCs are computed as a Fourier or cosine transform from the log-magnitude spectrum that is estimated through the frequency response of the all-pole filter defined by the prediction coefficients.

Perceptual linear prediction coefficients (PLPs): Based on the psychoacoustic principles such as critical band analysis (Bark), equal loudness pre-emphasis and intensity-loudness relationship, the PLPs is presentes. The PLP feature extraction as shown in figure 4.

Vocal tract features and voice source features

The traditional linear source-tract model can decompose the voice into two parts: the vocal tract and the voice source

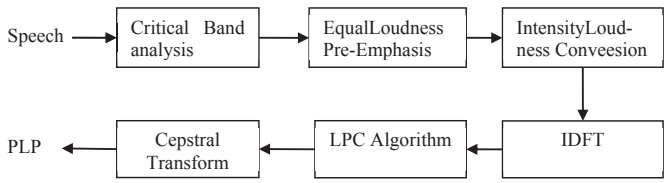

Figure 4: Overview of PLP feature extraction.

[50,51]. Auto-regressive moving average models corresponding to the formants and anti-formants are used to model vocal tract. The vocal tract features were extracted using extended Kalman smoothing, and these were used to analyzing depression [52,53]. The voice source was considered to carry speaker-specific information, since voice source features characterize the glottal excitation signal and fundamental frequency, and the glottal features are not directly measurable due to effect of the vocal tract filtering. It is assumed that the glottal source and the vocal tract are independent of each other.

Firstly, the linear prediction model is used to estimated the parameters of vocal tract, then the original waveform is inverse-filtered to obtain an estimate of source signal. There are various methods for inverse filtering have been implemented and evaluated, meanwhile, feature extraction methods based on the glottal flow estimate also have been proposed [54-56].

From the literature, voice source features are not as discriminative as vocal tract features, but the integration of these two complementary features can improve accuracy [57]. The literature [58], also shows that the voice source features need less training and testing data, which compared to the vocal tract features need the amount of data [59]. The reason is that vocal tract features depend on the phonetic content and thus sufficient phonetic coverage is required for both the training and test utterances. Voice source features, in turn, depend much less on phonetic factors.

Spectro-temporal features

The formant transitions and energy modulations of spectro-temporal signal contain effective speaker-specific information. A common way to contain some temporal information to features is through first-order and second-order time derivative estimates, known as delta (Δ) and double-delta (Δ^2) coefficients, respectively [60]. They are calculated as the time differences between the adjacent vectors feature coefficients and usually appended with the base coefficients on the frame level. The most commonly used is 13 MFCCs with Δ and Δ^2 coefficients, implying 39 features per frame. The spectro-temporal features were applied to SR, showing great performance when integrated with conventional MFCCs [61]. The energy is very easy to calculate. It defined as follows:

$$E_i = \log\left(\sum_{i=1}^{N_l} s_i(l)^2\right) \quad (5)$$

where the $s_i(l)$ and N_l respectively are the $i-h$ voice information on the flame l and the number of samples on the

Citation: Liu J, Philip Chen CL, Li T, Zuo Y, He P (2019) An overview of speaker recognition. Trends Comput Sci Inf Technol 4(1): 001-012.
DOI: http://dx.doi.org/10.17352/tcsit.000009

flame l. To calculate the Δ and Δ^2, the following formula is used:

$$\Delta = \frac{\sum_{i=1}^{N_l} i(x_{t+i} - x_{t-i})}{2\sum_{i=1}^{N_i} i^2} \quad (6)$$

where Δ is a delta coefficient, from frame t calculated in terms of the static coefficients x_{t+i} to x_{t-i}. A typical value for N is 2. Δ^2 coefficients are computed in the same way.

i-vector features

The i-vector is developed from the joint factor analysis (JFA) [62,63]. The i-vector can convert the variable length sequence of the original speaker's speech information to fixed dimension. In general, the super-vector M represent a speaker utterance, which obtained from the cascading of mean-vectors of all mixture components in the GMM. This speaker-dependent super-vector can be decomposed as:

$$M = m + Vy + Ux + Dz \quad (7)$$

where m is a speaker and session-independent supe-rvector, which generated from UBM. The matrix V and D both define the subspace of speaker, and U define a session sub-sapce. y and x represent the factors of speaker and channel respectively. Dz acts as a residual to compensate for the information of speaker that may not be caught by Vy.

The literature [64], found that the factors of channel also include speaker infromation, thus proposed a single subspace called total variability, which is known as the i-vector approch. The new speaker and session-dependent GMM super-vector is redefined as:

$$M = m + Tw \quad (8)$$

where T is a low rank matrix of speaker and session variability, and the total factor w is called identity vector, named i-vector [65].

Features based on deep neural networks

In 2006, deep learning is proposed by Hinton [66]. In recent years, DNNs has been successfully applied in speech recognition because of the the powerful feature extraction capability [67]. In the method for speaker recognition, DNN is often used to predict the speaker class for give a frame of speech. Deep speaker verification methods utilize various DNN structures to learn speaker features which are known as speaker representation and contain information of speaker that can be used for categorization. For Google's work, a DNN is trained to map frame-level features to the corresponding speaker identity target. During enrollment process, the average value of the last DNN hidden layer serves as the speaker model, which are defined as a deep vector or 'd-vector' [68]. In the evaluation phase, the decision was made by using the distance between the target d-vector and the test d-vector. Many researchers compared with the baseline i-vector model and replace the i-vector model [69,70] investigate the use of DNNs for a small foot-print TD speaker verification task, the trained DNN is used to extract speaker specific features from the last hidden layer. MIT Computer Science and Artificial Intelligence Laboratory investigate the use of DNNs to generated a stacked bottleneck (BN) feature representation for low-resourespeech recognition and the bottleneck is a small hidden layer placed in the middle of the network [71,72] described the development of a DNN bottlenenck feature i-vector system and demonstrated substantial performance gains [73]. Compares to the performance of deep locally-connected network (LCN) and convolutional neural network (CNN) for text-dependent speaker recognition. Speaker recognition is very sensitive to the duration of speech, and the recognition performance of short-time speech is a key point to decide whether it can be commercialized. Because of this, the x-vector emerged in this context. The DNN maps variable-length utterances to fixed-dimension embeddings, which were called x-vectors. The vectors are extracted from a TDNN and utilized like i-vectors. David et al. used data augmentation to improve performance of DNN embeddings for SR [74].

Sandro Cumani ect. proposed a speaker modeling approach which extracts a compact representation of a speech segment, similar to the speaker factors of JFA to i-vectors, refered to as "eigenvoice-vector", or "e-vector" for short. The process of estimating e-vector space is similar to i-vector training, and generates a more accurate speaker subspace [75]. The e-vector model is similar to the i-vector model:

$$s = m + Ew \quad (9)$$

where s and m are the GMM supervector and UBM mean supervector respectively and w is a d dimension random vector, with standard normal prior distribution. Since the goal is that E spans the same subspace of V, we define E as:

$$E = VA \quad (10)$$

where A is a full rank d × d matrix. Matrix E spans the same subspace of V because its columns are a linear combination of the columns of V. The specific calculation method reference [76].

Speaker Modeling

As above, the traing speech is passed through the front-end processing step (feature extraction) and the next step is the build a speaker model. There are many modeling techniques that have been utilized in speaker recognition systems. The choice of model is largely depended on the temporal techinique development level, the desired performance, easy to train and update, and the computation consideration. The methods of SR model can be grouped into three categories:

1. Template matching: The model contains a template which is a sequence of feature vectors from a fixed phrase. During verification, dynamic time warping (DTW) is used to align and measure the similarity between the test phrase and the speaker template to generate matching scores. This approach is

Citation: Liu J, Philip Chen CL, Li T, Zuo Y, He P (2019) An overview of speaker recognition. Trends Comput Sci Inf Technol 4(1): 001-012.
DOI: http://dx.doi.org/10.17352/tcsit.000009

utilized almost completely for text-dependent applications. In speaker recognition system, the most commonly used template maching methods are vector quantization (VQ).

2. **Probabilistic model:** In the training process, the effective feature vectoris extracted from one or more speakers and the corresponding mathematical model is estabilished according to its statistical characteristics. The model can describe the distribution of the speaker' feature vector. In the recognition process, the sequence of test speech is matched with the trained speaker model. The similarity between the test speech and the trained model is calculated from the perspective of probability statictics. In speaker recognition, the most commonly used models are Gaussian mixture models (GMMs), adapted Gaussian mixture models and hybrid model.

3. **Neural networks:** The specific model utilized in this technique can take many forms, the most commonly utilized in SR is the DNN, such as DCNN, LSTM, DBM etc. The main difference between the neural network model and other approaches described is that these models are trained to discriminate between the modeled speaker and some optional speakers.

Classical approaches

Vector quantization: The vector quantization is one of the simplest TI speaker models which is firstly proposed in the 1980s [77,78]. A set of short-term training feature-vectors for speaker can be utilized directly to represent the basic characteristics of the speaker. However, this direct representation is impractical when the number of training vectors is huge. Because the amount of memory and computation required become very large. Therefore, an effective method to compress training data is found by using vector quantization techniques. In this method, the codebooks of VQ include the numerical representation of features which are produced in the training phase by clustering the feature vectors of each speaker. We have utilized standard LBG algorithm to cluster a group of L training vectors into a group of M codebook vectors. In the recognition phase, each reference speaker' codebook is used to carry out vector quantization of input utterance, and VQ distortion accumulated in the whole input utterance is utilized for recognition and determination. The literature [79,80] offers competitive accuracy when compared with the probability model.

Gaussian Mixture Model (GMM): GMM is a stochastic model that can be deemed as the extend of the VQ model, in which the clusters are overlapping. x represents the feature-vector which have D-dimensional, the mixture density of speaker s is expressed as follow:

$$p\left(x|\lambda_s\right)=\sum_{i=1}^{M}p_i^s b_i^s\left(x\right) \tag{11}$$

The density is the weighted linear combination of M component single model Gaussian densities, $b_i^s\left(x\right)$, each parameterized by a mean vector, $u_i^s\left(x\right)$, and covariance matrix, Σ_i^s;

$$b_i^s\left(x\right)=\frac{1}{\left(2\pi\right)^{D/2}\left|\Sigma_i^s\right|^{1/2}}\times\exp\left\{-\frac{1}{2}\left(x-u_i^s\right)'\left(\Sigma_i^s\right)^{-1}\left(x-u_i^s\right)\right\} \tag{12}$$

The mixed weights p_i^s, the constraint conditions are furture satisfied $\sum_{i=1}^{M}p_i^s=1$, Collectively, the parameters of speakers density model are denoted as $\lambda_s=\left\{p_i^s,u_i^s,\Sigma_i^s\right\},i=1,\cdots M$. Maximum likelihood speaker model parameters are estimated using the iterative Espectation-Maximization (EM) algorithm [81]. In each EM iteration, the following reestimation formula is utilized to ensure the monotonic increase of model likelihood:

Mixture Weights: $p_i^s=\frac{1}{M}\sum_{i=1}^{M}p\left(i|x,\lambda\right)$ (13)

Weighted Means: $u_i^s=\frac{\sum_{i=1}^{M}p\left(i|x_i,\lambda\right)x_i}{\sum_{i=1}^{M}p\left(i|x_i,\lambda\right)}$ (14)

Variances: $\sigma_i^2=\frac{\sum_{i=1}^{M}p\left(i|x_i,\lambda\right)x_i^2}{\sum_{i=1}^{M}p\left(i|x_i,\lambda\right)}-u_i^2$ (15)

Gaussian mixture model is a stochastic model which has become the reference method in SR [82,83]. The literature [84], discussed the application of GMMs to speaker modeling in detail. The advantages of GMM is that has strong representation power for feature vector. Around 2000s, GMM-based systems were successly applied to the annual NIST speaker recognition evaluations (SRE). The GMM-based system has consistently produced state-of-art performace [85,86].

However, in turn, the parameter size also expands proportionately. Large scale variables make the training data unable to make GMM fully trained. Another disadvantage is that in higher-levels information about the speaker conveyed in the temporal speech signal are not utilized. Moreover, due to the relatively limited data that can be used in the speaker recognition system, when large amounts of parameters are estimated, there will be overfitting. Therefore, a more general model is proposed for speaker recognition.

Gaussian Mixture Model-Universal Backgroud Model (GMM-UBM): As discussed earlier, it is important to adapt acoustic models to new operating conditions because of data variability due to different speakers, environments, speaking styles and so on. The MIT Lincon Laboratory [87], adopting bayesian adaptation of speaker models from a universal backgroud model and handset-based score normalization. The system is refered to as the Gaussian Mixture Model-Universal Backgroud Model (GMM-UBM) [88] figure 5.

The whole GMM-UBM consists of three stages:

1) UBM training: In the TI speaker verification, training a UBM using a large number of non-target speakers. It can represent the distribution of general acoustic features of TI speaker recognition. The parameters are trained with the

Citation: Liu J, Philip Chen CL, Li T, Zuo Y, He P (2019) An overview of speaker recognition. Trends Comput Sci Inf Technol 4(1): 001-012.
DOI: http://dx.doi.org/10.17352/tcsit.000009

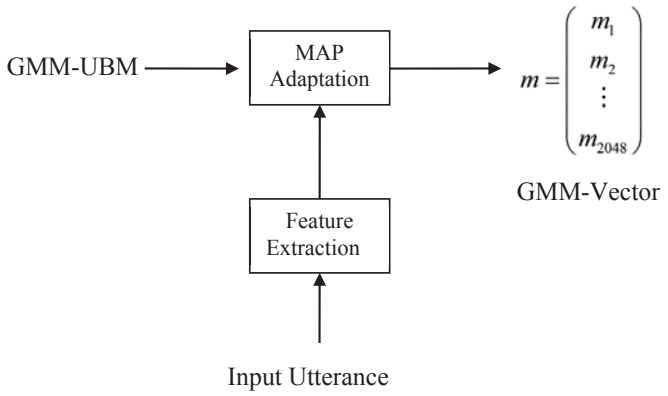

Figure 5: The Structure of GMM-UBM in Speaker recognition.

iterative EM algorithm and require unlabled data which contain different people, differernt text and different channels.

2) Enrollment stage: In this stage, the target speaker model is obtained by adaptive UBM parameters using the target speaker's enrollment speech and Maximum a Posteriori (MAP) adaptation. This adaptation would adjust the parameters of GMM mixtures which can be observed in the enrollment phase. The outputs of this stage are dependent on the speaker's models.

3) Verification stage: In this stage, the likelihood ratio decision method is utlized in the Verification phase. Suppose the O represents the extracted feature from the test utterance of speaker S, suppose that:

H_0: O is from the target speaker S

H_1: c is not from the target speaker S

The decision is made according to the likelihood ratio as follow:

$$\Lambda = \frac{1}{T}\log\frac{p(O|H_0)}{p(O|H_1)} = \begin{cases} \geq \theta & accept \quad H_0 \\ < \theta & accept \quad H_1 \end{cases} \tag{16}$$

where $p(O|H_i), i = 0,1$, is the probability of hypothesis H_i, which can be calculated using the function of probability density for O given the target speaker GMM model or the impostor GMM model. During the testing phase, the UBM model typically acts as an impostor model. T shows the number of frames in observation O.

The whole procedure of this method is easy to implement, the literature [89], obtain statisfactory performance in TD speaker verification and literature [90], proposed the text-independent speaker Verification system.

Inspired by the joint factor analysis theory, Dehak [91], proposed to extract a more compact vector called i-vector from mean supervector of GMM. The i-vector is equivalent to the identity of the speaker. As discussed earlier [92], presented a new speaker verification system where factor analysis is utilized to define a new low-dimensional space that models both speaker and channel variabilities [93]. Presented a straight-forward extension of the standard i-vector approach and show that the low-dimensional representation of utterances can be successfully used in TD speaker verification.

Deep neural network

In 2006, Hinton proposed deep learning aigorithms and deep neural networks. There is a wave of deep learning in industy and academia, and have achieved great success in the field of speech recognition and image processing. The success use of i-vectors in speaker recognition and deep learining techniques in speech processing applications has encouraged the research community to combine those techniques for speaker recognition. A possible use of deep learning techniques in speaker recognition is to combine them with the state-of-the-art i-vector approach. Many researchers compared to the baseline i-vector model and replace the i-vector model.

Two kinds of combination can be considered. Deep learning techniques can be used in the i-vector extraction process, or applied as a backend. Like the literature [94], extracted the feature from Deep Belief networks (DBN) and combined the MFCC features achieve better performance and [95], proposed an impostor selection algorithm and a universal model adaptation process in a hybrid system based on deep belief networks (DBN) and deep neural networks (DNN) to discriminatively model each target speaker [95]. Used DNN instead of the GMM to extract the i-vector. The result shows the DNN approach significantly improved the i-vector speaker recognition systen as compared to the traditional UBM-GMM approach [97]. Proposed the use of DNN senone (context-dependent triphones) posteriors for computing the soft alignments, which resulted in remarkable reductions in speaker recognition error rates [98]. Proposed a supervised GMM-UBM based on DNN posteriors, which have been successfully evaluated telephony speaker recognition. Speaker recognition based on deep neural network is the focus of current research and the trend of future research.

Decision Logic

In the speaker recognition, the general approach used in the speaker recognition system is to apply a likelihood ratio to determine if the claimed speaker is to be identtified. For an utterance $X = \{x_1, \cdots\cdots x_T\}$ and a claimed speaker identity with corresponding model λ_C, the likelihood ratio is

$$\frac{\Pr(X \text{ is from the claimed speaker})}{\Pr(X \text{ is't from the claimed speaker})} = \frac{\Pr(\lambda_c|X)}{\Pr(\lambda_{\bar{c}}|X)} \tag{17}$$

Applying Bayes' rule and discarding the constant prior probabilities for claimant and imposter speakers (they are accounted for in the decision threshold), the likelihood ratio in the log domain becomes

$$\Lambda(X) = \log p(X|\lambda_C) - \log p(X|\lambda_{\bar{C}}) \tag{18}$$

The term $p(X|\lambda_C)$ is the likelihood of the utterance given which is from the claimed speaker and $p(X|\lambda_{\bar{C}})$ is the likelihood

Citation: Liu J, Philip Chen CL, Li T, Zuo Y, He P (2019) An overview of speaker recognition. Trends Comput Sci Inf Technol 4(1): 001-012.
DOI: http://dx.doi.org/10.17352/tcsit.000009

of the utterance given that is not from the claimed speaker. The likelihood ratio is compared to a threshold θ.

$$\Lambda(X)\begin{cases}\geq\theta, & accept\\ \leq\theta, & reject\end{cases} \quad (19)$$

The likelihood ratio essentially measures how much better the claim's model scores for the test utterance, which compared to some non-claimed model. The decision threshold is then set to adjust the trade-off between rejecting true register utterances (false rejection errors) and accepting unregistered utterances (false acceptance errors).

The terms of the likelihood ratio are computed as follows. The likelihood of the utterance given the claimed speaker's model is directly computed as

$$\log p\left(X\left|\lambda_C\right.\right)=\frac{1}{T}\sum_{t=1}^{T}\log p\left(X_t\left|\lambda_{\bar{C}}\right.\right) \quad (20)$$

the $\frac{1}{T}$ scale is used to normalize the likelihood for utterance duration.

After 2012, the mainstream algorithm is from statistics-based machine learning to deep learning, Deep neural network (DNN)-based bottleneck feature, d-vectors, x-vectors, and j-vectors tend to appear successively. There are two approaches generally to solving decision: the one for similarity, the other for back-end classifier. There are two methods to calculating similarity, the first method is cosine similarity, the larger the value is, the higher the similarity will be, as adopted by Baidu Deep Speaker. Another method is Euclidean distance, the smaller the value is, the higher the similarity will be, as adopted by Facenet. The back-end classifier is used to produce recognition scores by comparing vectors from different utterances. The most commonly used back-end classifier is SVM and probabilistic linear discriminant analysis(PLDA).

End-to-end modeling

In previous studies, the speaker recognition is divided into two parts, one is the front-end feature extraction (the part 4), the other is the back-end speaker modelling (the part 5) and dicision-making (the part 6). But since 2006, K. Q. Weinberger etc. proposed the concept of triplet loss, which is a kind of loss function in deep learing, and is used to training samples with small differents. The training data includes anchor sample, positive sample and negative sample. In order to achieve the similarity calculation of sample by optimazing the distance between the anchor sample and the positive sample is smaller than the distance between the anchor sample and the negative sample [99]. The triplet loss was first applied in face recognition [100]. Inspired by the advancements in triplet loss for FaceNet, the end-to-end model is proposed and first applied in speaker recognition.

In 2016, Georg Heigold etc. proposed an end-to-end text-dependent speaker verificaiton system, which applied a Deep neural network (DNN) with a locally-connected layer followed by fully-connected layer as the baseline DNN and a long short-term memory recurrent neural network (LSTM) with a single output [101]. The architecture is optimized using the end-to-end loss:

$$l_{e2e}=-\log p(\text{target}) \quad (21)$$

where

$\text{target}\in\{accept, reject\}$,

$p(accept)=\left(1+\exp\left(-wS\left(X, spk\right)-b\right)\right)^{-1}$ and

$p(reject)=1-p(accept)$. The end-to-end system will result in better accuracy without the need for heruistics and postpocessing steps. Based on the literature [102], and the end-to-end system proposed by G. Heigold and researches from Google Inc., David Snyder etc. proposed an end-to-end text-independent speaker verification system. The research showed that neural network-based end-to-end system is genetally applicable to verification tasks and the end-to-end outperformed i-vector model [103]. Herv´e Bredin firstly applied triplet loss to speaker turn embedding, which also first used Long Short-Term Memory recurrent networks in speaker turn embedding [104]. Microsoft Coporation is inspired by recent advancements in both face recognition and speaker recognition, Chunlei Zhang etc. present a novel end-to-end text-independent speaker verification system, which is referred to as Inception-resnet-v1 network and apply the triplet loss function to optimize the entire system [105]. Based on previous research, Chunlei Zhang etc. continue to investigate end-to-end text-independent speaker verification by incorporating the variability from different utterance durations. They modify the previous network architecture which only can encode fixed length features. The new netwok is referred to as Inception-resnet-v1 with Spatial Pyramid Pooling [106]. Chao Li etc. researches from Baidu Inc present Deep Speaker, which is a nueral speaker embedding system that mapps utterances to a hypersphere where speaker similarity is measured by cosine similarity. The embeddings generated by Deep Speaker can be used for many tasks, including speaker identification, verification, and clustering [107]. Li Wan of Google proposed the Generalized End-to-End loss function to train speaker verification models more efficiently which makes the training of speaker verification model more efficient than triple loss-based end-to-end loss function [108,109].

Conclusion and Future Research Trends

This paper presents a systematic overview on various features extraction methods and back-end models for speaker recognition. In particular, we summarized the method of deep learning in speaker recognition. It also gave an overview of the latest end-to-end methods in speaker recognition. Through this reviews, we have observed that the performance significantly depends on the features which represent the speaker specific characteristics and modelling technique, and also depends on the nature of the speech signal, speech

Citation: Liu J, Philip Chen CL, Li T, Zuo Y, He P (2019) An overview of speaker recognition. Trends Comput Sci Inf Technol 4(1): 001-012.
DOI: http://dx.doi.org/10.17352/tcsit.000009

duration and phonetic environment. For SR system, the most important part is the front-end feature extraction. However, in all the studies presented above, there is no generalized universal feature extraction approach, and different speech durations have a great influence on the final decision. Although deep learning has greatly improved the accurancy of speaker recognition, even though effective, deep learning system are highly data-driven and massive amounts of data are needed for training the background models. The data sets need to be labeled and organized in the requriments of actual model. If the development data conditions do not match to the expected operation environment, the training accuracy will drop significantly. It is clear that we can not spend a lot of time on data tagging in actual research. Much of the recent process in speaker recognition largely rely on supervised learning, there has been very limited work on unsupervised speaker recognition approaches, which involve creating speaker models based on feature extracted from unlabeled data. The types of feature and feature extraction methods are great challenges in future research.

In particular, we provided a review of end-to-end techniques for speaker recognition, which combined the front-end and back-end. Compared with the traditional DNN-based i-vector approach, end-to-end model significantly improve the speaker recognition system. But the model size of end-to-end system is very large, and the requirements of CPU are also high. This will lead to end-to-end model sufferring from the time-consuming training process since a great mumber of hyperparameters and complicated structures are involved. Hence, the future work will focus on reducing model size, reducing CPU requirements and also look for ways of improving the long training time.

References

1. Jain AK, Ross A, Prabhakar S (2004) An introduction to biometric recognition. IEEE Transactions on Circuits and Systems for Video Technology 14: 4-20. Link: http://bit.ly/30CQzL5

2. Bhardwaj S, Srivastava S, Hanmandlu M, Gupta JR (2013) GFM-based methods for speaker identification. IEEE Trans Cyber 43: 1047-1058. Link: http://bit.ly/30CQXcv

3. Apsingekar VR, De Leon PL (2009) Speaker model clustering for efficient speaker identification in large population applications. IEEE Trans on Audio, Speech, and Language Processing 17: 848-853. Link: http://bit.ly/2HtDjRi

4. Schmidt L, Sharifi M, Lopez-Moreno I (2018) Large-scale speaker identification dataset. arXiv:1706.08612v2 [cs.SD].

5. Togneri R, Pullella D (2011) An overview of speaker identification: Accuracy and robustness issues. IEEE Circuits and Systems Magazine 11: 23-61. Link: http://bit.ly/31ZIn7Q

6. Kinnunen T, Li H (2010) An overview of text-independent speaker recognition: From features to supervectors. Speech Communication 52: 12-40. Link: http://bit.ly/2HrX9MU

7. Reynolds DA (2002) An overview of automatic speaker recognition technology. IEEE. Link: http://bit.ly/2Nwp3uS

8. Gish H, Schmidt M (1994) Text-independent speaker identification. IEEE Signal Processing Magazine 11: 18-32. Link: http://bit.ly/2L8LJPc

9. Bimbot F, Bonastre JF, Fredouille C, Gravier G, Chagnolleau IM, et al. (2004) A tutorial on text-independent speaker verification. EURASIP Journal on Applied Signal Processing 4: 430-451. Link: http://bit.ly/3235ONC

10. Chen K, Wang L, Chi H (1997) Methods of combining multiple classifiers with different features and their applications to text-independent speaker identification. International Journal of Pattern Recognition and Artificial Intelligence 11: 417-445. Link: http://bit.ly/31ZJjcm

11. Wan V, Campbell WM (2000) Support vector machines for speaker verification and identification. Neural Networks for Signal Processing X. Proceedings of the 2000 IEEE Signal Processing Society Workshop (Cat. No.00TH8501). Link: http://bit.ly/2zmZWCq

12. Assaleh KT, Mammone RJ (1994) New LP-derived features for speaker identification. IEEE Transactions on Speech and Audio Processing 2: 630-638. Link: http://bit.ly/2La4BNQ

13. Chandran V, Ning D, Sridharan S (2004) Speaker identification using higher order spectral phase features and their effectiveness vis–a–vis Mel-Cepstral features. International Conference on Biometric Authentication 614-622. Link: http://bit.ly/2La8sKH

14. Atal BS (1974) Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. J Acoust Soc Am 55: 1304-1322. Link: http://bit.ly/2ZmI9uE

15. Doddington GR (1935) Speaker Recognition- Identifying People by their Voices. Proceedings of the IEEE 73: 1651-1664. Link: http://bit.ly/2Pd4lmd

16. Reynolds DA (1995) Speaker identification and verification using Gaussian mixture speaker models. Speech Communication 17: 91-108. Link: http://bit.ly/2ZunCzz

17. Wang JC, Chin YH, Hsieh WC, Lin CH, Chen YR, et al. (2015) Speaker Identification with Whispered Speech for the Access Control System. IEEE Transactions on Automation Science and Engineering 12: 1191-1199. Link: http://bit.ly/2zpv0BS

18. Shahin I (2017) Speaker Identification in the Shouted Environment Using Suprasegmental Hidden Markov Models. Link: http://bit.ly/2ZrU7T7

19. Kerst LG (1962) Voiceprint Identification. The Journal of the Acoustical Society of America 34: 725. Link: http://bit.ly/2ZsgT9e

20. LUCK JE (1969) Automatic speaker verification using cepstral measurements. The Journal of the Acoustical Society of America 46: 1026. Link: http://bit.ly/2HsrwCG

21. Atal BS (1976) Automatic recognition of speakers from their voices. Proceedings of the IEEE 64: 460-475. Link: http://bit.ly/2U6pHAv

22. Young S (2000) Woodland PState clustering in HMM-based continuous speech recognition. Computer Speech and Language 70-79.

23. Saraclar Z, Nock H, Khudanpur S (2000) Pronunciation modeling by sharing Gaussian densities across phonetic models. Computer Speech and Language 14: 137-160. Link: http://bit.ly/2PhPtTL

24. Reynolds D (1996) MIT Lincoln Laboratory site presentation. in Speaker Recognition Workshop, A Martin, Ed, sect 5, Maritime Institute of Technology, Linthicum Heights, MD 27-28.

25. Davis S, Mermelstein P (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Transactions on Acoustics, Speech, and Signal Processing 28: 357-366. Link: http://bit.ly/30CFf1w

26. Reynolds DA (1992) A Gaussian mixturemodeling approach to textindependent speaker identification. Ph.D. thesis, Georgia Institute of Technology, Atlanta, Ga, USA. Link: https://b.gatech.edu/2U5Upd6

27. Campbell JP (1997) Speaker recognition: A tutorial. Proceedings of the IEEE 85: 1437-1462. Link: http://bit.ly/328eW3D

Citation: Liu J, Philip Chen CL, Li T, Zuo Y, He P (2019) An overview of speaker recognition. Trends Comput Sci Inf Technol 4(1): 001-012.
DOI: http://dx.doi.org/10.17352/tcsit.000009

28. Martin A, Przybocki M (2000) The NIST 1999 Speaker Recognition Evaluation - An Overview. Digital Signal Processing 10: 1-18. Link: http://bit.ly/2PctZYr

29. Sturim D, Campbell W, Dehak N, Karam Z, McCree A, et al. (2011) The MIT LL 2010 speaker recognition evaluation system: Scalable language-independent speaker recognition. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Link: http://bit.ly/345pwKH

30. Scheffer N, Ferrer L, Graciarena M, Kajarekar S, Shriberg E, et al. (2011) The SRI NIST 2010 speaker recognition evaluation system. ICASSP. Link: http://bit.ly/2Pjh7zM

31. Reynolds DA, Quatieri TF, Dunn RB (2000) Speaker verification using adapted Gaussian mixture models. Digital Signal Processing 10: 19-41. Link: http://bit.ly/2zpVm6K

32. Campbell WM, Sturim DE, Reynolds DA (2006) Support vector machines using GMM supervectors for speaker verification. IEEE Signal Processing Letters 13: 308-311. Link: http://bit.ly/341M8LO

33. Bimbot F, Bonastre JF, Fredouille C, Gravier G, Chagnolleau IM, et al. (2004) A tutorial on text-independent speaker verification. EURASIP Journal on Applied Signal Processing 4: 430-451. Link: http://bit.ly/3235ONC

34. Schmidt M, Gish H (1996) Speaker identification via support vector classifiers. in Proc IEEE Int Conf Acoustics, Speech, Signal Processing. Link: http://bit.ly/2PhRMpT

35. Gu Y, Thomas T (2001) A text-independent speaker verification system using support vector machines classifier. in Proc. European Conference on Speech Communication and Technology.

36. Kenny P (2006) Joint factor analysis of speaker and session variability: theory and algorithms. Draft Version. Link: http://bit.ly/31ZOLfk

37. Dehak N, Dehak R, Kenny P, Brümmer N, Ouellet P, et al. (2009) Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. Interspeech. Link: http://bit.ly/344P1f8

38. Campbell WM, Sturim DE, Reynolds DA, Solomonoff A (2006) SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. IEEE International Conference on Acoustics Speech and Signal Processing Proceedings. Link: http://bit.ly/2Zgb2Zs

39. Lei Y, Scheffer N, Ferrer L, McLaren M (2014) Novel Scheme For Speaker Recognition Using A Phonetically-Aware Deep Neural Network. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Link: http://bit.ly/2PeqRex

40. Richardson F, Reynolds D, Dehak N (2015) Deep Neural Network Approaches to Speaker and Language Recognition. IEEE Signal Processing Letters 22: 1671-1675. Link: http://bit.ly/33Y8jm8

41. Variani E, Lei X, McDermott E, Moreno IL, Dominguez JG (2014) Deep Neural Networks for Small Footprint Text-Dependent Speaker Verification. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Link: http://bit.ly/2NBySI6

42. Chen N, Qian Y, Yu K (2015) Multi-Task Learning for Text-dependent Speaker Verification. Interspeech. Link: http://bit.ly/2HrKVUD

43. Snyder D, Romero DG, Sell G, Povey D, Khudanpur S (2018) X-Vectors: Robust Dnn Embeddings For Speaker Recognition. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) Link: http://bit.ly/2Zxakq1

44. Heigold G, Moreno I, Bengio S, Shazeer N (2015) End-to-End Text-Dependent Speaker Verification. Link: https://tinyurl.com/yy49seyc

45. Reynolds D, Andrews W, Campbell J, Navratil J, Peskin B, et al. (2003) The SuperSID project: exploiting high-level information for high-accuracy speaker recognition. IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings. Link: http://bit.ly/30DyAnH

46. Tirumala SS, Shahamiri SR, Garhwal AS, Wang R (2017) Speaker identification features extraction methods: A systematic review. Expert Systems with Applications 90: 250-271. Link: http://bit.ly/2ZgMiQN

47. Davis SB, Mermelstein P (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing 28: 357-366. Link: http://bit.ly/30CFf1w

48. Makhoul J (1975) Linear prediction: A tutorial review. Proceedings of the IEEE 63: 561-580. Link: http://bit.ly/2zpYOOK

49. Hermansky H (1990) Perceptual linear predictive (PLP) analysis of speech. The Journal of the Acoustical Society of America 87: 1738. Link: http://bit.ly/2zkFSAS

50. Fant G (1960) Acoustic Theory of Speech Production. The Hague, the Netherlands, Mouton. Link: http://bit.ly/2MECvgI

51. Wong DY, Markel JD, Gray JAH (1979) Least squares glottal inverse filtering from the acoustic speech waveform. IEEE Transactions on Acoustics, Speech, and Signal Processing 27: 350-355. Link: http://bit.ly/2U59JH3

52. Mehta DD1, Zañartu M, Feng SW, Cheyne HA, Hillman RE (2012) Mobile voice health monitoring using a wearable accelerometer sensor and a smartphone platform. IEEE Trans Biomed Eng 59: 3090-3096. Link: http://bit.ly/2ZteuPK

53. Williamson JR, Quatieri TF, Helfer BS, Ciccarelli G, Mehta DD (2014) Vocal and facial biomarkers of depression based on motor incoordination and timing. 4th International Audio/Visual Emotion Challenge and Workshop: Depression Challenge 65-72. Link: http://bit.ly/2Zs87fA

54. Plumpe MD, Quatieri TF, Reynolds DA (1999) Modeling of the glottal flow derivative waveform with application to speaker identification. IEEE Transactions on Speech and Audio Processing 7: 569-586. Link: http://bit.ly/2ZDhMAd

55. Gudnason J, Brookes M (2008) Voice Source cepstrum coefficients for speaker identification. IEEE International Conference on Acoustics, Speech and Signal Processing. Link: http://bit.ly/2Zw3Esc

56. Gudnason J, Thomas MRP, Ellis DPW, Naylor PA (2012) Data riven voice source waveform analysis and synthesis. Speech Communication 54: 199-211. Link: http://bit.ly/329ycxQ

57. Murty K, Yegnanarayana B (2006) Combining evidence from residual phase and MFCC features for speaker recognition. IEEE Signal Processing Letters 13: 52-55. Link: http://bit.ly/324MhMJ

58. Chan W, Zheng N, Lee T (2007) Discrimination power of vocal source and vocal tract related features for speaker segmentation. IEEE Transactions on Audio, Speech, and Language Processing 15: 1884-1892. Link: http://bit.ly/2NB8eyY

59. Prasanna SRM, Gupta SC, Yegnanarayana B (2006) Extraction of speakerspecific excitation information from linear prediction residual of speech. Speech Communication 48: 1243-1261. Link: http://bit.ly/30DFKID

60. Furui S (1981) Cepstral analysis technique for automatic speaker verification. IEEE Transactions on Acoustics Speech and Signal Processing 29: 254-272. Link: http://bit.ly/341Wsna

61. Rabiner L, Juang BH (1993) Fundamentals of Speech Recognition. Link: http://bit.ly/327FvGc

62. Kenny P, Boulianne G, Ouellet P, Dumouchel P (2007) Joint factor analysis versus eigenchannels in speaker recognition. IEEE Transactions on Audio, Speech, and Language Processing 15: 1435-1447. Link: http://bit.ly/2HvgXz4

63. Kenny P, Boulianne G, Ouellet P, Dumouchel P (2007) Speaker and session variability in gmm-based speaker verification. IEEE Transactions on Audio, Speech, and Language Processing 15. Link: http://bit.ly/2Zs9fzQ

Citation: Liu J, Philip Chen CL, Li T, Zuo Y, He P (2019) An overview of speaker recognition. Trends Comput Sci Inf Technol 4(1): 001-012.
DOI: http://dx.doi.org/10.17352/tcsit.000009

64. Dehak N (2009) Discriminative and generative approaches for long- and short-term speaker characteristics modeling: application to speaker verification. Link: http://bit.ly/2MDE1zO

65. Dehak N, Kenny P, Dehak R, Dumouchel P, Ouellet P (2011) Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech, and Language Processing 19: 788-798. Link: http://bit.ly/343UXoi

66. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521: 436-444. Link: http://bit.ly/2Pf726Y

67. Hinton G, Deng L, Yu D, Dahl G, Mohamed AR, et al. (2012) Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Processing Magazine 82-97 Link: https://tinyurl.com/y328hjpr

68. Variani E, Lei X, McDermott E, Moreno IL, Dominguez JG (2014) Deep Neural Networks for Small Footprint Text-Dependent Speaker Verification. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Link: https://tinyurl.com/y3bdfv6m

69. Lei Y, Scheffer N, Ferrer L, McLaren M (2014) A Novel Scheme For Speaker Recognition Using A Phonetically-Aware Deep Neural Network. ICASSP. Link: http://bit.ly/2KYepLQ

70. Chen YH, Moreno IL, Sainath TN, Visontai MO, Alvarez R, et al. (2015) Locally-Connected and Convolutional Neural Networks for Small Footprint Speaker Recognition. Interspeech Link: https://tinyurl.com/y5ghwwr4

71. Zhang Y, Chuangsuwanich E, Glass J (2014) Extracting Deep Neural Network Bottleneck Features Using Low-Rank Matrix Factorization. IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP). Link: http://bit.ly/2Zho4WF

72. Richardson F, Reynolds D, Dehak N (2015) Deep Neural Network Approaches to Speaker and Language Recognition. Ieee Signal Processing Letters. 22: 1671-1675. Link: http://bit.ly/2ZmEufN

73. Chen YH, Moreno IL, Sainath TN, Visontai MO, Alvarez R, et al. (2015) Locally-Connected and Convolutional Neural Networks for Small Footprint. Interspeech Link: https://tinyurl.com/y5ghwwr4

74. Snyder D, Romero DG, Povey D, Khudanpur S (2017) Deep Neural Network Embeddings for Text-Independent Speaker Verification. Interspeech. Link: http://bit.ly/2HsH09M

75. Cumani S, Laface P (2018) Speaker Recognition Using e-Vectors. IEEE /ACM Transactions on Audio, Speech, and Language Processing 26: 736-748. Link: http://bit.ly/30AzyBi

76. Cumani S, Laface P (2018) Speaker Recognition Using e-Vectors. IEEE /ACM Transactions on Audio, Speech, and Language Processing 26: 736-748. Link: http://bit.ly/30AzyBi

77. Burton D (1987) Text-dependent speaker verification using vector quantization source coding. Handbook of Biometrics 151-170. Link: http://bit.ly/2PjCbWS

78. Soong FK, Rosenberg AE, Juang BH, Rabiner LR (2012) A vector quantization approach to speaker recognition. Proceedings of International Conference on Innovation & Research in Technology for Sustainable Developmen

79. Yu K, Mason J, Oglesby J (1995) Speaker recogntion using hidden Markov models dynamic time warping and vector quantisation", IEEE Proceedings Vision, Image and Signal Processing 142: 313-318. Link: http://bit.ly/2Zknesc

80. Desai D, Joshi M (2014) Speaker Recognition Using MFCC and Hybrid Model of VQ and GMM. Recent Advances in Intelligent Informatics 235: 53-63. Link: http://bit.ly/341BdBX

81. Dempster AP, Laird NM, Rubin DB (19771) Maximum likelihood from incomplete data via the EM algorithm. 39: 1-38. Link: http://bit.ly/326IT3W

82. Reynolds D, Rose R, (1995) Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Transactions on Speech and Audio Processing 3: 72-83. Link: http://bit.ly/2ZwtKeO

83. Reynolds DA (1995) Speaker identification and verification using Gaussian mixture speaker models. Speech Communication 17: 91-108. Link: http://bit.ly/2ZunCzz

84. Reynolds DA (1992) A Gaussian mixturemodeling approach to textindependent speaker identification. PhD thesis, Georgia Institute of Technology, Atlanta, Ga, USA. Link: https://b.gatech.edu/2U5Upd6

85. Doddington G, Przybocki M, Martin A, Reynolds DA (2000) The NIST speaker recognition evaluation—overview, methodology, systems, results, perspective. Speech Communication 31: 2-3. Link: http://bit.ly/30y6qdP

86. Martin A, Przybocki M (2000) The NIST 1999 Speaker Recognition Evaluation - An Overview. Digital Signal Processing 10: 1-18. Link: http://bit.ly/2PctZYr

87. Reynolds DA (1997) Comparison of background normalization methods for text-independent speaker verification. Eurospeech Link: http://bit.ly/2U5Ur4Q

88. Reynolds DA, Quatieri TF, Dunn RB (2000) Speaker verification using adapted Gaussian mixture models. Digital Signal Processing 10: 19-41. Link: http://bit.ly/2zpVm6K

89. Sturim DE, Reynolds DA, Dunn RB, Quatieri TF (2002) Speaker verification using text-constrained gaussian mixture models. IEEE International Conference on Acoustics, Speech, and Signal Processing Link: http://bit.ly/2U82Jt5

90. Liu Y, Fu T, Fan Y, Qian Y, Yu K (2014) Speaker verification with deep features. International Joint Conference on Neural Networks (IJCNN) Link: http://bit.ly/2Ht0MSD

91. Dehak N (2009) Discriminative and generative approaches for long- and short-term speaker characteristics modeling: application to speaker verification. Link: http://bit.ly/2MDE1zO

92. Dehak N, Kenny P, Dehak R, Dumouchel P, Ouellet P (2011) Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech, and Language Processing 19: 788-798. Link: http://bit.ly/343UXoi

93. Zeinali H, Sameti H, Burget L (2017) HMM-Based Phrase-Independent i-Vector Extractor for Text-Dependent Speaker Verification. IEEE/ACM Transactions on Audio, Speech, and Language Processing 25 Link: http://bit.ly/2U6JIMV

94. Ali H, Tran SN Benetos E, d'Avila Garcez AS (2018) Speaker recognition with hybrid features from a deep belief network. Neural Computing and Applications 29: 13-19. Link: http://bit.ly/2ZgP2hb

95. Ghahabi O, Hernando J (2017) Deep Learning Backend for Single and Multisession i-Vector Speaker Recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing Link: http://bit.ly/2Zlo1cj

96. Lei Y, Scheffer N, Ferrer L, McLaren M (2014) A Novel Scheme For Speaker Recognition Using A Phonetically-Aware Deep Neural Network. ICASSP. Link: http://bit.ly/2KYepLQ

97. Lei Y, Scheffer N, Ferrer L, McLaren M (2014) A Novel Scheme For Speaker Recognition Using A Phonetically-Aware Deep Neural Network. ICASSP. Link: http://bit.ly/2KYepLQ

98. Snyder D, Romero DG, Povey D (2015) Time delay deep neural network-based universal background models for speaker recognition. IEEE Workshop on Automatic Speech Recognition and Understanding Link: http://bit.ly/2NyngFO

99. Weinberger KQ, Blitzer J, Saul LK (2006) Distance metric learning for large margin nearest neighbor classification. In NIPS. MIT

Citation: Liu J, Philip Chen CL, Li T, Zuo Y, He P (2019) An overview of speaker recognition. Trends Comput Sci Inf Technol 4(1): 001-012.
DOI: http://dx.doi.org/10.17352/tcsit.000009

100. Schroff F, Kalenichenko D, Philbin J (2015) FaceNet: A Unified Embedding for Face Recognition and Clustering. Link: http://bit.ly/2zm9LAz

101. Heigold G, Moreno I, Bengio S, Shazeer N (2016) End-to-End Text-Dependent Speaker Verification. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) Link: http://bit.ly/2LcVV9g

102. Variani E, Lei X, McDermott E, Moreno IL, Dominguez JG (2014) Deep Neural Networks for Small Footprint Text-Dependent Speaker Verification. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Link: https://tinyurl.com/y3bdfv6m

103. Snyder D, Romero DG, Sell G, Povey D, Khudanpur S (2016) Deep neural network-based speaker embeddings for end-to-end speaker verification. in Spoken Language Technology Workshop (SLT) Link: http://bit.ly/2KX0ENA

104. Bredin H (2016) Tristounet: Triplet Loss for Speaker Turn Embedding, arXiv:1609.04301v2 [cs.SD] Link: http://bit.ly/323sgGo

105. Zhang C, Koishida K (2017) End-to-End Text-Independent Speaker Verification with Triplet Loss on Short Utterances. Interspeech Link: http://bit.ly/2MI0Guz

106. Zhang C, Koishida K (2017) End-To-End Text-Independent Speaker Verification with Flexibility in Utterance Duration. ASRU Link: http://bit.ly/2L0mrUF

107. Li C, Ma X, Jiang B, Li X, Zhang X, et al. (2017) Deep Speaker: an End-to-End Neural Speaker Embedding System. arXiv:1705.02304v1 [cs.CL] Link: http://bit.ly/2ZgSngd

108. Wan L, Wang Q, Papir A, Moreno IL (2017) Generalized End-To-End Loss for Speaker Verification. arXiv:1710.10467v1 [eess.AS] Link: http://bit.ly/2Hslt1k

109. Philip Chen CL, Liu Z (2018) Broad Learning System: An Effective and Efficient Incremental Learning System Without the Need for Deep Architecture. IEEE Transactions on Neural Networks and Learning Systems 29: 10-24. Link: http://bit.ly/30CvjFp

Citation: Liu J, Philip Chen CL, Li T, Zuo Y, He P (2019) An overview of speaker recognition. Trends Comput Sci Inf Technol 4(1): 001-012.
DOI: http://dx.doi.org/10.17352/tcsit.000009