Distant Speaker Recognition An Overview

Article *in* International Journal of Humanoid Robotics · July 2015

DOI: 10.1142/S0219843615500322

CITATIONS
27

READS
1,368

1 author:

Mohammad Ali Nematollahi
Islamic Azad University
79 PUBLICATIONS **517** CITATIONS

SEE PROFILE

International Journal of Humanoid Robotics
Vol. 12 (2015) 1550032 (45 pages)

DOI: 10.1142/S0219843615500322

Distant Speaker Recognition: An Overview

Mohammad Ali Nematollahi* and S. A. R. Al-Haddad
Department of Computer & Communication Systems Engineering, Faculty of Engineering, University Putra Malaysia, UPM Serdang, 43400 Selangor Darul Ehsan, Malaysia
**greencomputinguae@gmail.com*

Received 20 September 2013
Accepted 11 May 2015
Published 22 July 2015

Distant speaker recognition (DSR) system assumes the microphones are far away from the speaker's mouth. Also, the position of microphones can vary. Furthermore, various challenges and limitation in terms of coloration, ambient noise and reverberation can bring some difficulties for recognition of the speaker. Although, applying speech enhancement techniques can attenuate speech distortion components, it may remove speaker-specific information and increase the processing time in real-time application. Currently, many efforts have been investigated to develop DSR for commercial viable systems. In this paper, state-of-the-art techniques in DSR such as robust feature extraction, feature normalization, robust speaker modeling, model compensation, dereverberation and score normalization are discussed to overcome the speech degradation components i.e., reverberation and ambient noise. Performance results on DSR show that whenever speaker to microphone distant increases, recognition rates decreases and equal error rate (EER) increases. Finally, the paper concludes that applying robust feature and robust speaker model varying lesser with distant, can improve the DSR performance.

Keywords: Distant speaker recognition; far-field speaker recognition; reverberation; ambient noise; robust speaker recognition.

1. Introduction

Many biometric features such as face, iris, hand, signature, fingerprint and voice have been applied for biometric systems. However, human voice is more interesting for developers where the speech signal has valuable information on who/what/how speaker speaks.[1,2] As it is easy to be produced, captured, transmitted and it is noninvasive, it does not need direct contact or being perceived by the individual. It also has lower cost in comparison with other recognition techniques due to easy, little and cheap hardware requirements.[3] Due to these reasons, a speech recognition system is applied for human robot interaction (HRI).[4,5] It must be noted that many HRI tools have been equipped with microphone which capture speech signals from the environment.[6] Basically, robotic auditory system is implemented based on three

main functions including source localization, source separation, and source recognition.[7] However, distant speaker recognition (DSR) as a part of source recognition is more concerned in robotics systems due to interaction purposes. Therefore, high speech quality is required to increase DSR performance. Many issues such as background noise, echo, reverberation, microphone type, microphone location, talker direction, environmental and Lombard effect can affect DSR performance.[8–12] Due to these challenges, more additional steps such as channel selection, speaker tracking and localization, blind source separation (BSS), beamforming and feature enhancement should be added to conventional speaker recognition systems for a DSR system integration.

The DSR can be divided into speaker identification, which involves identifying an unknown speaker by using known speakers population, and speaker verification as the most popular part of general biometric verification method[1] which aims to verify the identity of a given speaker from the population of known speakers.

The demand for this application comes from various fields, namely, tele-commerce, automobile industry, robotics, forensics, airports, smart home and office environments, and law courts. In other words, the application would be useful wherever there is a need to recognize speaker conversation over voice channel such as telephone or wireless phone or voice over IP (VoIP).

In this tutorial paper, fundamental researches on DSR[9,10,12–15] are reviewed. Although for sakes of brevity basic information can be referred to a valuable book[8] and reviewed papers,[16–19] this paper is dedicated to state-of-the-art of DSR techniques, problems and solutions. Various parts in DSR system of pre-processing, robust feature extraction, feature normalization, robust speaker modeling, model compensation, recognition, score normalization, decision making and different metrics for evaluation will be discussed. We have tried to concentrate on new features and robust models for DSR technology. Also, model compensation, feature normalization techniques and dereverberation which have been used properly in DSR systems will be discussed to open research opportunity for future trends in this field. Some performance results on recognition rates and equal error rate (EER) on DSR will also be discussed. Finally, conclusion and future trends are discussed.

This paper is organized as following: Section 2 discuses speech production system in close and far distant. Furthermore, various challenges and problems are discussed. Section 3 discussed major steps for pre-processing of the distant speech to optimized time, storage and performance. Sections 4 and 5 present different robust acoustic feature extraction and feature normalization for improving the DSR performance. Section 6 presents how features are modeled by statistical speaker model. Section 7 shows how these models are robust against possible mismatch between training and testing sets. Sections 8 and 9 explain speaker recognition (scoring) and score normalization techniques. The related state-of-art of DSR to humanoid robots are deeply contextualizing in Sec. 10. Various metrics for DSR system evaluation are explained in Sec. 11. Performance of different DSR systems are compared in terms of evaluation metrics by presenting some results in Sec. 12. Dereverberation techniques

and well-known DSR databases are presented in Sec. 13. Finally, conclusion and future work are discussed in Sec. 14.

2. Speech Signal and DSR

Speech is generated when air is exhaled from the lungs which passes through the throat, vocal cords, mouth and nasal tract which can be modeled as in "Eq. (1)"

$$S(f) = G(z) \cdot H(z) \cdot R(z), \tag{1}$$

where $S(f)$ corresponding to original speech signal, $G(z)$ is modeled as impulse train for voiced or white noise for unvoiced and $R(z)$ is modeled as a fixed differentiator[20] as "Eq. (2)"

$$R(z) = 1 - \alpha z^{-1} \quad (0.9 < \alpha < 1) \tag{2}$$

and $H(z)$ is vocal tract system which will be discussed in Sec. 3.

Generally, close speech is captured at a distant less than 0.16 ft to the speaker mouth. However, in distant situation, the microphone records speech at a distant more than 1 ft from the speaker source. Signal to noise ratio (SNR) in close speech is higher than distant speech due to lesser background noise and reverberation. Nevertheless, when the speaker (who is producing the speech signal) to microphone distant increases, the SNR of the speech (which is produced by speaker) decreases. The distant speech can be formulated[14] as "Eq. (3)"

$$\underbrace{Y_d[n]}_{\text{Distance component}} = \overbrace{X_d[n]}^{\text{Direct component}} + \underbrace{\sum_{i=1}^{m} \alpha_i X_d[n - n_i]}_{\text{Reverberant component}} + \overbrace{\vartheta[n]}^{\text{Ambient noise}}, \tag{3}$$

where $Yd[n]$ is recorded signal at distant of d, $Xd[n]$ is the direct component or original speech signal at time of n at distant of d, α and $\vartheta[n]$ are the gain factor and the background noise, respectively. For close recording, the reverberation and noise are tending to zero then $Yd[n] \approx Xd[n]$.

In Automatic speaker recognition (ASR), it is assumed that microphone is at a fixed point and close to the speaker (called, close speech) so it has lesser challenges and problems and immunes the speech signal in term of degradation factors such as ambient noise and reverberation. That's why, it is difficult to make ASR commercially viable. For real and practical situation like HRI and intelligent room environments, it is better to have distant or hands-free microphone from the speaker's mouth. Two main problems in DSR which cause the mismatch between training and test data includes the distortion (environmental issues), speaking style and accent. Currently, many techniques are available to compensate the mismatch due to intra-speaker and inter-speaker differences by articulation effects and higher-level linguistic features to capture speaker's idiosyncrasies.[21] Whenever high speech quality has been applied to speaker recognizer, the system performance has improved due to

decrease in mismatch between training and testing sets. That's why a lot of researches have been done on DSR to enhance speech quality by various techniques including beamforming, speech enhancement among others.[8,22] When a speaker produces speech, various undesired distortion factors affect the speech signal quality before the sounds can be recorded by the microphone. In the following, the definitions of these environmental distortion factors are described as

- Coloration: Certain frequency of speech would be amplified due to enclosed environment (where surround or close off on all sides) capacity.
- Head orientation and radiation which changes the direction of wave's sound pressure level (SPL) for microphone.
- Ambient noise (background noise) is any other additive signal (except desired speech signal) which contaminates desired speech signal. The noise is divided into stationary, which its statistical parameters are constant over time, and nonstationary which changes its statistical parameters in a short period of time.
- Echo and reverberation: While the speech signal travels the direct path between source to microphone, other part of the speech signal may arrive after a short period of time (for example; this part may travel the distant between source to wall and then wall to microphone). This short period of time (called delay) must be more than 0.1 s to be heard by human.

Generally, the recorded speech by microphone is classified into three parts: First, direct path when the speech travels the distant from source to microphone; second is early reflection when the speech travels indirect pass and reach the microphone between 50–100 ms after direct wave recorded by microphone; lastly, late reflection which may appear like noise because the wave may pass large indirect distant and fading its energy. Here energy is inversely proportional to time square.[23] In contrast to ambient noise, echo and reverberation are correlated with desired speech signal and can be expressed as "Eq. (4)"

$$Y(n) = h(n) * x(n) = \sum_{m=0}^{M} h(n)x(n-m). \tag{4}$$

Figure 1 presents availability of the ambient noise, reverberation and direct speech in distant speech recording. As seen, different microphones can record different amount of the speech signal due to spatial orientation. Moreover, time delay (τ) between recorded speeches can be applied for source-to-microphone distant estimation, speaker localization and detection.[8,24,25] Currently, many robots apply two omnidirectional microphones which their position are equipped as the human ears for recording sound. This method is called binaural. Binaural can overcome distant speech problems[26,27] because human can suppress these effects by binaural nature of hearing. In the following, a section is dedicated to discuss about problems and solution in binaural DSR systems.

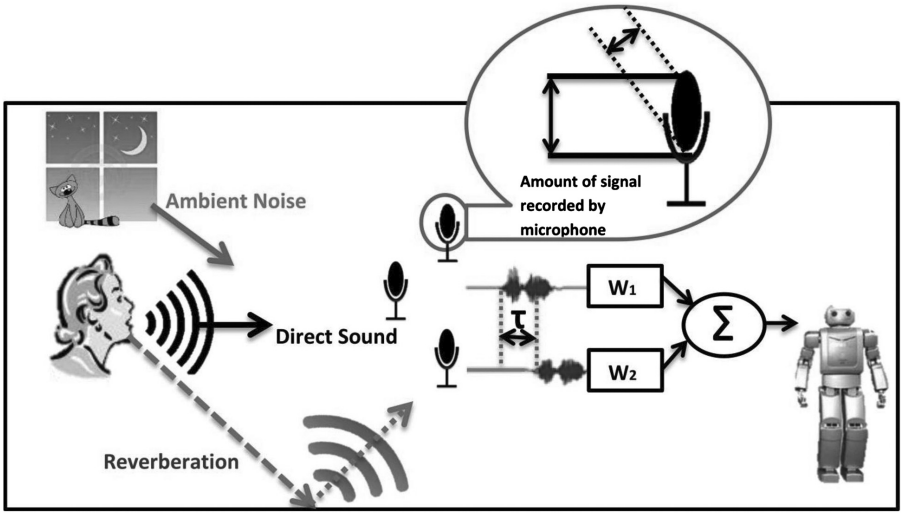

Fig. 1. Speech signal components.

Generally, spectral and physiological characteristics of speech including pitch, tone and volume is produced by speech patterns which themselves depend on various movements of the lips, tongue and palate (called the articulators). Speech is a biometric signal even for twins[28] due to various unique features like size and shape of the mouth and vocal and nasal tract, vocal cord's tension, and the combination of the properties imposed by the organs entangled with human speech. Moreover, behavioral or long term properties like prosody can be used for recognition as well as the physiological properties or short term features. Prosody involves variation in syllable length intonation, format frequencies, pitch, rate and rhythm of speech. All these features varied from one speaker to another.

Another speech characteristic is determined by formants frequency (also known as resonant frequency denoted as F1, F2 and F3) which are spectral parameters as shown in Fig. 2. Formants frequencies are described by the physical configuration of the vocal tract. Although different formants frequencies are available in speech signal especially for vowels, for speaker/speech recognition only three of them are used (F1, F2 and F3). Relative frequency locations of the formants are different from one person to another. That is why it is considered as a biometric feature too. Generally for spectral parameter estimation, 20–30 ms segments of the speech signal are used as stationary where shape of the vocal tract varies relatively slowly. Stress pattern of speech signal (like prosody, manifests spectral trajectory and distribution of energy in the spectrogram) also can be considered as biometric feature. This type of feature is called "high level" feature because it is estimated by observing multiple segment of speech signal.[29,30]

Number of reflections in a surrounded room can also be approximately estimated by calculating the ratio between the sphere volume (with radius microphone to

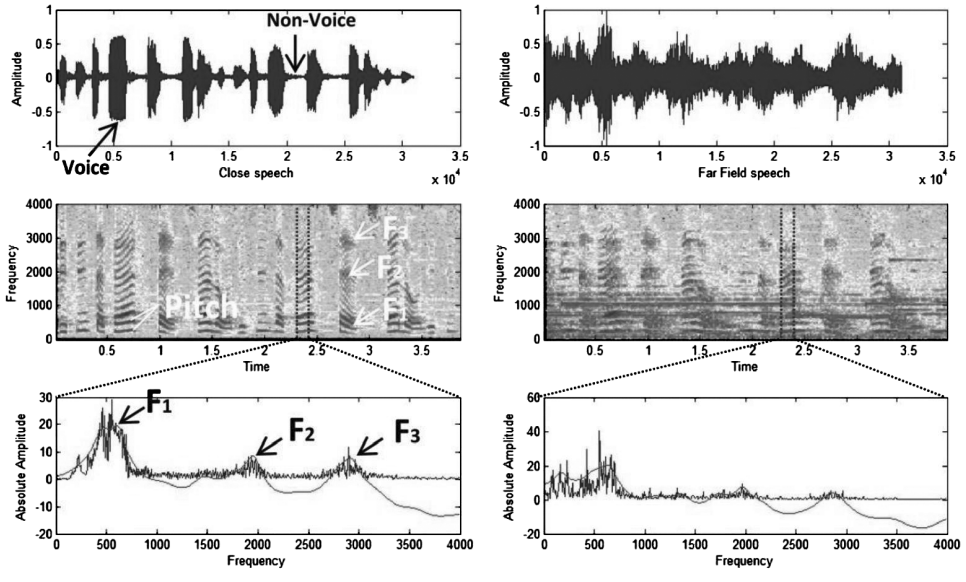

Fig. 2. Time-frequency representation of same speech signal at (a) close and (b) distant through the spectrogram.

distance) and room volume.[23] Before presenting the simple example on noise and reverberation in distant speech recording, spectral visualization technique, called spectrogram, must be described. Spectrogram applies short time fourier transform (STFT) or wavelet transform (WT) over a short period of time where trends of speech are quasi-stationary due to shape of the vocal tract which changes slowly over time. Spectrogram displays the frequencies of vocal cords vibration (pitch), and speech signal amplitude (volume) as they vary with time or some other variables. Two main spectrograms are wide-band spectrogram which has short window and provides good time resolution, but smears the harmonic structure and narrow-band spectrogram which uses long window and provides good frequency resolution but poor time resolution.

A speech signal which is simultaneously recorded in close distance and at a far distance is shown in Fig. 2. As seen, background noise and reverberation affect the speaker-specific information (such as format frequency and pitch) which are crucial for speaker recognition process. In the next section, some current techniques are presented to bypass, combat and limit these types of distortions.

Figure 3 presents different phases in a DSR system. In DSR when distant speech is recorded, it is pre-processed to attenuate distant distortion effect. Depending on the system, many techniques such as speech enhancements and pre-quantization (PQ) may be applied on distant speech. However, framing and windowing are mandatory process in DSR. Then, feature extraction phase is applied on pre-processed speech to

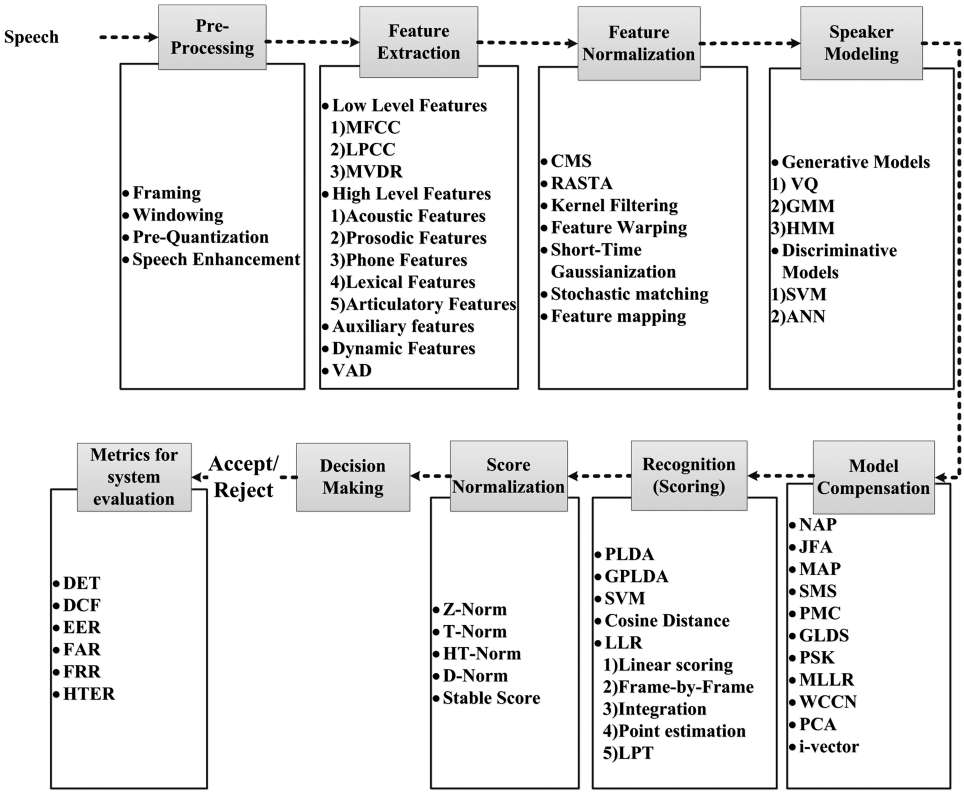

Fig. 3. Overall phases in DSR system.

extract speaker biometric features. This features are related to physical (low-level features) or behavioral (high-level) structure of speakers. Furthermore, auxiliary and dynamic features may be extracted to improve the DSR performance. Voice activity detection (VAD) is also applied to remove the silent part of the speech signal because a few features are available in the silence part. After feature extraction phase, feature normalization phase is applied to filter the distortions which is contaminated the extracted features. These phases can improve the DSR performance due to the mismatch compensation between extracted training and testing features. The most effective feature normalization techniques include cepstral mean subtraction (CMS), RelAtiveSpecTrAl (RASTA), kernel filtering, short time gaussianization (STG), stochastic matching and feature warping.

The normalized features are fed to speaker modeling phases to generate a pattern for each speaker. The speaker modeling is based on statistical techniques such as generative model (vector quantization (VQ), gaussian mixture model (GMM) and hidden Markova model (HMM)) and discriminative models (support vector machine (SVM) artificial neural network (ANN)). Due to large distortion in DSR system, speaker model should be adjusted by model compensation phase. This phase

improves the DSR performance by updating the speaker's model parameters in a more discriminative way. The most effective model compensation techniques are including nuisance attribute project (NAP), joint factor analysis (JFA), maximum *a posteriori* (MAP), speaker model synthesis (SMS), parallel model combination (PMC), generalized linear discriminative sequence (GLDS), probabilistic sequence kernel (PSK), maximum likelihood linear regression (MLLR), within-class covariance normalization (WCCN), principle component analysis (PCA) and identity-vector (i-vector).

In recognition phase, test utterance is scored based on training speaker model. Feature normalization and model compensation cannot normalize speaker-dependent score offset. Therefore, score normalization phase is used on DSR to normalize the effect of channel variability. The most effective score normalization techniques are zero normalization (Znorm), test normalization (Tnorm), hand-set test normalization (HTnorm), distant normalization (Dnorm) and stable score. In the decision making phase, normalized score is compared to a threshold to distinguish between claimant and impostors speakers in speaker verification; or determine the matching level in speaker identification application. Finally, DSR performance is evaluated by using some metrics such as decision error trade-off (DET), detection cost function (DCF), EER, false acceptance rate (FAR), false rejection rate (FRR), half total error rate (HTER) and recognition rate.

3. Pre-Processing of DSR

Basically, pre-processing is important for DSR systems before its feature extraction. However, the step of pre-processing is different from one DSR system to another. Generally, three steps will be performed on speech signal before feature is extracted from it. First, pre-emphasis filter is applied as in (2) to flat speech spectral for lesser finite precision effects. Second, speech signal is segmented to frames with size of 20–30 ms. Finally, windowing is applied to each frame to reduce discontinuities at start and end of the frames.

Recently, many PQ techniques have been proposed to reduce the number of frames before feature extraction.[31] PQ is based on assumption that the amount of speaker-specific information vary from a frame to another. Due to gradual movements of articulation in speech signal in practice, the adjacent frames have similar features. However, the main PQ condition is that the speaker recognition system without using PQ techniques have to be larger time than speaker recognition systems with PQ. The PQ techniques not only must reduce the cost and complexity of the speaker recognition system but also they must improve the recognition system.

These PQ techniques are proposed to remove vector redundancy by using distant measurement, probability density function (PDF), randomization, averaging, decimation and clustering.[32,33] In PQ distant measurement technique, the distant between vectors are computed based on two method including probabilistic

(Fisher ratio) and nonprobabilistic (Euclidian, weighted Euclidian) distant measurement. In Euclidian and weighted Euclidian distance, similarity between vectors are measured. However, in Fisher ratio distance, dissimilarity between vectors is given. In PQ PDF technique, higher order statistic (HOS) has been applied to select vectors. The main idea in PDF PQ technique is that, HOS is more affected by larger samples of speech which are likely sourced of mismatch between training and testing sets. In randomization PQ technique, vectors is selected randomly. In averaging PQ technique, mean of the vectors is used. In decimation PQ technique, every Nth vector of the speech signal is selected for feature extraction. Finally in clustering PQ technique, the speech sequence is partitioned in N cluster by using Linde–Buzo–Gray (LBG) clustering technique. PQ clustering technique apply pruning methods such as static pruning, hierarchical pruning, adaptive pruning and confidence-based pruning which are different in pruning criteria.

Speech enhancement techniques can be applied in DSR as pre-processing step to attenuate the ambient noise and reverberation. They remove the noise based on statistical estimation of noise. Speech enhancement has classified as direct and indirect techniques.[8] The direct speech enhancement includes Wiener filter, wavelet thresholding, Gaussian and super-Gaussian MMSE estimation and stereo-based piecewise linear compensation for environments (SPLICE). However, in indirect speech enhancement, the distorted speech signal is enhanced by four steps: First step is distortion estimation which estimates the probability density of distortion; second step is distortion evolution which predicts/updates the distortion for next frame due to nonstationary nature of distortion. The third step is distortion evaluation which evaluates the predicted/updated distortion based on clean speech signal. The final step is distortion compensation which applies subtraction or inverse filtering to derive the features of estimated clean signal involved in compensation process.[8]

Many speech enhancement techniques assume noise and reverberation as separate problems; however, for achieving the full potential, speech enhancement techniques should be able to filter these types of distortions only in single step effectively. Some researchers have been trying to model noise and reverberation in high dimensional space by applying particle filter.[34] There is a tradeoff between single-channel and multi-channel speech enhancement. Although multi-channel can be applied for denosing and dereverberation, it is not useful in look direction when direct component of desired speaker is captured directly. This degradation can be enhanced by using adopting objective functions in single channel speech enhancement before or after beamforming. On the other hand, HOS can be incorporated into novel beamforming and post-filtering algorithms to improve the performance of DSR systems. It is crucial to stop the distinction between post-filtering and beamforming by offering comprehensive solutions.

Applying speech enhancement technique has two major drawback. First, it increases the processing time especially in real-time application. Second, it may remove the speaker-specific information as well as noise.

4. Robust Feature Extraction

Feature modeling requires speech signal to transform to dimension-reduced vectors in acquiring specific information on biometric features during recognition.[29,30] This is applied after digitalization, pre-emphases and windowing process. Information can be acquired based on two main levels of features, namely, "Low-level" and "High-level"[29,30] and a generalized form of low-level which is known as auxiliary.[35] Low-level features reveal physical structure of a given speaker including vocal cords, excitation and pitch, while high-level features detect the speech's behavior, e.g., prosody, phonetic, and conversational patterns. Low-level and high-level features are also used in different time duration for feature extraction. Short frames of speech (< 30 ms) must be applied for extracting low-level features because physical features are assumed to be stationary (not changing) in a short duration of time. Nevertheless, more than a few second time-scales are needed to extract high-level features properly. Due to low complexity and low overhead, many of speaker recognition systems use low-level features.[13,36–38] The feature fusion might be able to combine the auxiliary features as well as other low-level features.

4.1. *Low-level features*

The most popular algorithms of low-level features extractions for speaker recognition are perceptual linear prediction cepstral (PLPC), linear prediction cepstral coefficients (LPCC) and mel frequency cepstral coefficients (MFCC). In contrast to MFCC features that apply FFT power spectrum, the LPCC and PLPC apply smoothed spectrum by using an all-pole model. Here, the LPCC is described due to its effectiveness for ASR.[36,39]

On the other hand, applying these features for the DSR systems causes mismatch conditions between training and testing phases. Lower robustness in these features is because of the distant distortion factors such as reverberation and ambient noise. For these reasons, many researchers investigate robust variants of MFCCs or novel feature extraction algorithm, such as mel scale logarithmic spectrum (MSLS), linear frequency cepstral coefficient (LFCC), human factor cepstral coefficients (HFCC), equivalent rectangular bandwidth (ERB) exponential frequency cepstral coefficients (EFCC), bark scale frequency cepstral coefficients (BFCC) and minimum variance distortion less response (MVDR).[17,40,41] Also, fundamental frequency variation (FFV) can be applied as complimentary features with MFCC and MVDR under mismatch situation to improve the system performance.[11,42] Although the MFCC shows better results for clean condition without significant differences between training data, the PLPC can give better results for noisy or mismatched condition. It is demonstrated that the ERB scale works more robustly than the Bark and the Mel scale for distant microphone recording and using the BFCC features is not better than using MFCC features.[43] On the other hand, the warped or perceptual MVDR spectral envelope provides better results for close speech as well as distant recording.[44]

Neither DFT nor LP cannot provide robust spectral envelope in presence of noise.[45] Therefore, several robust LP techniques are proposed to tackle problems caused by ambient noise. These techniques include weighted linear prediction (WLP), stabilized WLP (SWLP), and regularized LP technique (RLP, RWLP and RSWLP). Instead of speech enhancement, WLP increases the contribution of the samples which are less corrupted by noise to optimist the all-pole filter. WLP tries to minimize the energy of the predication error. However, WLP does not guarantee stability of the all-pole filter and it can decrease DSR performance.[45] Therefore, SWLP solves stability of the WLP by choosing weight function properly. This weight function is chosen based on short time energy (STE) window with length of M Selecting small M emphasizes samples with larger value which is caused due to less vulnerable to noise.[46] Furthermore, regularized LP techniques have smoothed spectra without affecting position of the formants. This technique can solve the problems of formants over-sharping by using double autocorrelation lag.[47] Regularized LP techniques show larger improvement for speaker recognition performance than other spectral techniques under noise.

4.2. *Auxiliary features*

The auxiliary features sometimes refer to voice source features.[48] The feature fusion might combine the auxiliary features as well as other low-level features. They contain other information like Rover posteriors,[35] non-nativeness score and phoneme as compared to low level features and improve the performance of the system. In some other applications, the residual speech signal found in LPC is used for estimating the glottal flow waveform.[49–51]

4.3. *Dynamic features*

Under some circumstances like Lombard effect[8] or high ambient noise, the low level features might not work well. That explains why the temporal information feature can improve the performance.[52] The most well-known temporal features include: (a) average of first-order dynamic derivation known as features velocity (denoted as Δ features); (b) average second-order dynamic derivation known as acceleration of features (denoted as $\Delta\Delta$ features). Also, the energy features like Logarithm of the total energy of feature (L2 norm); and its first-order temporal derivative can be applied for recognition purpose.

4.4. *Voice activity detector (VAD)*

With the exception of silence and speech located in voice signal, most of the features are located in speech part. VAD or Speaker Activity Detection or speech activity detection (SAD) is utilized in DSR systems to distinguish speech and silence segments in speech signal. Thus, it helps to improve the recognition task. In some

situations, due to background noise or other environmental disturbances, it is difficult to distinguish silence and speech segments. Thus, some parameters including energy level, periodicity, zero crossing rate (ZCR), magnitude sum function (MSF), long-term spectral divergence (LTSD), and HOS are useful for VAD in DSR systems.[53] Even though these techniques have been used for VAD, none of them are suitable and robust especially for binaural robot case due to differences and dynamic volume changes which are caused by ambient noise and reverberation. Basically, VAD is combined with speaker localization and recognition in a robotic systems.

4.5. *High-level features*

High-level features can play major roles in different ways in state of the arts for DSR due to four important issues. First, they can be a bridge between the gap of text-dependent and text-independent DSR because the low-level features are properly conditioned based on high-level features. Consequently, the low-level features are transformed to features with less variance from distortion and large inter-speaker variability which leads to better speaker discrimination and improved performance. Second, these high-level features are speaker-specific which have more tolerance to distortion. Third, some speaker specific information (nonuniform extraction region features (NERFs)) is beyond the low-level features (frame level) e.g., prosody, which subsumes energy, pitch and duration. Fourth, in some situations, the DSR should generalize and model different internal recognizer features. There are many high-level features but some researchers have used Cepstral or other low-level features to find high-level information e.g., phonemic. Some popular high-level features are including: Acoustic feature, prosodic features, phone features, lexical features, and articulatory features.

4.6. *Speaker recognition versus speech recognition*

Generally, speech recognition concentrates on speech linguistic information without considering the speaker, whereas speaker recognition focuses on speaker-specific information by ignoring phonemic and linguistic information. In contrast to speech recognition which normalizes source of speech variability like speaker dependency, in speaker recognition, the main aim is normalizing speaker's features variability such as phone, words, pronunciation and so on. For example, the frequency bands for phonetic discrimination in speech recognition are within the range of low and mid frequency bands (200 Hz to 3 kHz).[54] For speaker recognition, discriminative speaker features are within low and high frequency bands (for glottis is between 100 to 400 Hz, for piriform fossa is between 4 and 5 kHz, and for constriction of the consonants is 7.5 kHz).[55] Although MFCC is not extracted speaker-specific information exclusively, it often considers as a standard for comparison due to simplicity, generality and conventionality.[9,10,12,13,16,17,19,21,26,27,29–31,33,49,50,52,53,56–100]

5. Feature Normalization

Sometimes, the mismatch between trained and real environment can influence performance. Many environmental effects such as reverberation, cross-channel effect, noise and room impulse response (RIR) corrupt distant speech recording. As a result, the extracted features are distorted. Currently, three techniques are applied to compensate the adverse effect on a mismatch between training and testing condition for DSR[12,83]: (1) feature based compensation (normalization) technique which takes place before the features are fed to recognizer and try to compensate the noisy features for better matching between training and testing sets. (2) Model base adaption technique which uses noisy data to train/adapt the speaker model (for more details refer to Sec. 7). (3) Score normalization which removes score shifts and scaling caused by the mismatch conditions (for more details refer to Sec. 9).

Feature normalization techniques has divided in two classes such as model-based (CMN, mean and variance normalization (MVN), short-time mean and variance normalization (STMVN), short-time mean, and scale normalization (STMSN) techniques); and data distribution-based (STG and histogram normalization) techniques.[101] In model-based normalization technique, certain statistical properties of speech such as mean, variance, moments, are normalized to reduce the residual mismatch in feature vectors. However, data distribution-based techniques aim at normalizing the feature distribution to the reference.

5.1. *Cepstral mean normalization (CMS)*

Average of the MFCC is computed for all frames of speech. It is then deducted from each of the coefficients; similar to "Homeomorphic" filtering. In other words, channel variation is shown as offsets in MFCC coefficients. In some situations the variance of the MFCC coefficients is normalized to enhance the robustness of Cepstral features. The main disadvantage of this technique is its inefficiency for additive white noise. This method is successfully applied in series of DSR systems.[13,22] Also, the spectral subtraction (SS) cannot be efficient for low SNR.[102] A new CMS technique called position dependent cepstral mean normalization (PDCMN)[84] is proposed for DSR systems that can be proper for short utterance and real-time processing. PDCMN is estimated from utterances recorded *a priori* depending on the environment.

5.2. *RASTA filtering*

It is a type of CMS used by speaker recognition systems to eliminate the cross channel conflicts by using band-pass filtering for attenuation too slowly or quickly Cepstral components which are out of typical range of change in speech signal. In Ref. 103, two tests are conducted by changing the speaker position, namely, same-position and cross-position tests. It can be concluded that a few amount of speaker

related information is in Cepstral mean and many information is outside of the RASTA filtering pass-band. Also, the CMN can enhance the performance in many situations and degrades performance in fewer situations.

5.3. *Kernel filtering*

Low-level features just extract linear features of the speech signal; however, they might not extract nonlinear features properly. The majority of nonlinear features remain constant even under channel noise, so they are more robust than linear features. Nonlinear dynamic models, e.g., nonlinear maximum likelihood feature transformation, nonlinear transformation/mapping, kernel based time-series features, nonlinear discriminant techniques, neural predictive coding and other auxiliary methods can extract much nonlinear features. These nonlinear features are including vowels (which is different from speaker to speaker) and nonchaotic behavior (which is obtained by low dimensional manifolds with order smaller than four).

A new technique, called "Kernel predictive coefficients" (KPCs), can extract nonlinear features from speech signal by applying nonlinear filtering methods of a functional regression procedure in a reproducing Kernel Hilbert space (RKHS). This technique does not take any hypothesis on channel (static or dynamic) into account and it is known as semi-parametric procedure.

Kernel filtering approach for robust speaker recognition is perceived. It suppresses noise to preserve high order speaker-specific information for better discrimination by applying nonlinear smoothing method. It starts by windowing the speech signal and applying regression, which is nonlinear function. The result is then mapped to low dimensional manifold by nonlinear function. After that, DCT is applied on manifold parameters to obtain KPC coefficients. Even if the power increases through the noise, the KPC spectra are not changed. By applying different projection methods on this features, different KPC are obtained.

5.4. *Feature warping*

The observed Cepstral feature distribution is mapped to a normal distribution by using sliding window. This technique has more computational overhead due to its construction method which whitens the Cepstral feature's distribution and causes normal distribution for each speech frame. Nevertheless, it can establish very robust Cepstral features.

5.5. *STG*

It is similar to feature warping with just one difference. Just apply linear transformation before mapping features to a normal distribution. Linear transformation can be computed by expectation-maximization (EM) algorithm which makes the resulting features better suited to diagonal covariance GMMs.

5.6. *Stochastic matching*

It calculates the covariance matrix of test and train data, and then by applying a linear transform, the test data is scaled, rotated and translated to train data for finding overlapping between train and test data. The parameters of linear transformation would be adapted/re-estimated, if the test data have mismatched with a true speaker. This technique also increases the mismatch between the test and train data when they belong to different speakers.

5.7. *Feature mapping*

It reduces the channel variability by transforming the special channel's characteristics to an independent space for channel characteristics called "Normalization by supervision". This method is computed through recognition step, by mapping, adapting, and detecting the GMMs dependent features to channel independent features space.

6. Robust Speaker Modeling

When the features are extracted, it is fed to train/enrollment of a speaker-specific model. In the training phase, the feature vectors of known and unknown speakers are modeled by pattern matching techniques, then, these models will be used in recognition phase. Pattern matching can be classified as stochastic (parametric) models and template (nonparametric) models. Stochastic modeling apply probabilistic to an unknown and fixed PDF. The parameters of the PDF for training data are estimated by likelihood of the observation which is the given model. For evaluating, the likelihood of test utterance with respect to the trained model is estimated as pattern matching. However, in template model, pattern matching is deterministic and it is measured by comparing training and testing sets directly to find degree of similarity between them.

Dynamic time warping (DTW) and hidden Markov model (HMM) are examples of text-dependent speaker recognition for template and stochastic model respectively, and VQ and GMMs are examples of text-independent speaker recognition for template and stochastic model respectively. Generally, DTW, HMM, SVM, and VQ are applied for speaker verification. However, speaker identification applies GMM and ANNs.

Speaker modeling is a process of generating a pattern or model for each speaker inside the trained speaker's database. Depending on the speech type, the expected performance, the ease of training and updating, and storage and computation considerations, speaker model would be selected. Generally, speaker models in term of training paradigm can be classified into two distinct categories: Generative and discriminative.

6.1. *Generative models*

The objective of generative training is to enable the model to properly capture the empirical PDF of the acoustic speech which corresponds with the feature vectors. It

can be applied to GMMs or HMM. GMM is a single-state of HMM where the PDF is a mixture of Gaussians. Moreover, GMM has static pattern opposed to HMM with a sequential patterns. Furthermore, GMM may be considered as an extension of the VQ model, in which the clusters overlap. Instead of the VQ that assigned nearest cluster to feature vectors, the GMM applies a nonzero probability for each cluster during origination.[17]

6.1.1. *VQ-based modeling*

Some robot systems benefits from VQ technique for speaker recognition[2,99,103] due to low computational cost, the extensive variability of the speech signals and efficient codebook updating.

6.1.2. *HMM-based modeling*

HMM is a statistical model which captures both the temporal (dynamic) evolution of feature sequences and the statistical variation of the features.[103] HMM is not generally applied for DSR systems due to a significant increase in computational complexity.

6.1.3. *GMM-based modeling*

GMM can train faster in comparison with other models. Furthermore, it can be simply updated and scaled to add a new speaker. A GMM model composed of a finite mixture of multivariate Gaussian components which estimates a general PDF of model. In another type of GMM, named as universal background model (UBM), also well known as speaker independent world model, large number of speech data from large number of imposter speakers are used for training. GMM is widely used for speaker recognition systems due to text-Independent characteristic, probabilistic framework (which is more robust), computationally efficient, and easy to be implemented. However, GMM degrades significantly under distant speech recording due to mismatch between training and testing sets. In Ref. 10, a new method is applied for DSR based on various compensation techniques and high level features. In this method, rather than single GMM model for each speaker, multiple GMMs are trained for each speaker by multiple channel data. Then, the highest log likelihood score of all GMM models is chosen to be the frame score. In Ref. 103, the author concluded that the optimum order of GMMs must be determined based on training data and SNR. Even as the SNR is increased, the GMM order should be increased too.

6.2. *Discriminative models*

The objective of discriminative model is to estimate the discrimination of the target speaker from imposter and to minimize the error on a set of genuine and impostor training samples. It can be achieved by SVM or ANN.

6.2.1. *SVM-based modeling*

SVM is mathematical model which constructs the training set with few data point. SVMs utilize super-vectors as the input for recognition and training. Super-vector is a single vector which shows an entire speech utterance. It is generally created by mean vector of trained GMM. A system is proposed by combining two GMM and two SVM subsystems which each subsystem is trained for clean and contaminant conditions. The results show that DSR performance in term of EER for GMM (2.54%) and SVM (4.15%) can be improved to 1.91% when both GMM and SVM are combined.[104] The ability of learning can be provided for trading-off between its complexity and performance by SVM training's regularizer. Not only SVM is proper complimentary methods for speaker modeling, but also provide better performance and complexity as compare to ANN. Also, using SVM decision tree as a layered feature selection method may improve the performance.[4]

6.2.2. *ANN-based modeling*

ANN is the most powerful discrimination model which is applied in DSR systems when statistical distributions of the system are not known in advance. ANN needs less parameters and has higher performance compared to VQ model. ANN can learn and compensate the multipath distortion by mapping Cepstrum coefficients of distant speech to the close-talking speech and, thereby, enhances the performance.[90] Multi-layer network (MLP) and deep neural networks (DNNs) have been applied to extract a nonlinear feature transformation, which is called Bottleneck feature. By using unsupervised pre training DNN on conventional MFCC, the identification performance is improved around 46.3% in reverberant condition for distant talking speaker identification systems.[105]

Although many robot systems[88,106] apply ANN as their speaker modeling,[107,108] show that both GMM and ANN systems have similar performance when same microphone is applied for training and testing and the GMM appears better than ANN when different microphone is applied. However, it seems that GMM is outperformed than ANN totally.[56] Despite the strengths of ANN to discriminate features, there are three main drawbacks of applying this method: (1) Meeting the optimal state or configuration is not easily reachable. (2) Lots of data are needed for validation and training purposes. (3) Large training time and more expensive techniques.

6.3. *System fusion*

Fusion, combining information from different source of evidence, has been widely applied to improve the DSR performance. Two types of fusion such as decision fusion and feature fusion can be useful. Depending on the type of information that will be combined, different data fusion techniques are applied. For instance, if the model output is probabilistic, linear or log opinion pools is applied. If the model output is actually class labels, voting or ranking can be used. However, Fuzzy technique can

also be applied in combination. In Ref. 93 data fusion for speaker recognition is classified into Fuzzy integral method, Voting/ranking methods, Dempster–Shafer approach, Log opinion pool and Linear opinion pool.

Most of these systems assume that the features are ideal, i.e., different features are independent of each other. By using this assumption, different classifiers concentrate on different area of discrimination boundary. A hybrid model obtained from combination of a low-level feature such as pitch, or Cepstrum with a high-level feature such as conversational patterns and prosody. It must be mentioned that some DSR techniques[13] apply hybrid speaker model to improve the performance.

7. Speaker Model Compensation

Speaker model compensation techniques are trying to modify/update the parameters of the speaker model in order to learn distortions characteristics to improve performance. Although channel based distortion is adjusted for model parameters to represent test data in distant speech recognition, it is not applicable for DSR due to the accompanying loss of speaker discrimination information.

Various compensation techniques have been proposed for discrimination and generative speaker models. Factor analysis (FA) method, for instance, is proposed for GMMs based recognition that is based on random manner of the GMM.

There are three main steps (component) in FA based on GMM. (1) Speeches are captured from distant (Eigen-channel), (2) super-vector, also known as GMMs means, is calibrated and sorted before comparison process (Eigen-voice), and finally, (3) the channel variability is compensated by using JFA (classical MAP).[109] JFA regards the variability of GMMs super-vector based on linear equation ("Eq. (5)").

$$M = S + C. \tag{5}$$

Since M is speaker and channel dependent, it can be decomposed to speaker S and Channel C. S and C are statistical independent super-vectors. Channel variability can be modeled as $\boldsymbol{C} = \boldsymbol{UX}$ as well, in which U is Eigen-channel which is estimated from given data set and X is channel factor estimated from the speech utterance. X has been estimated based on speaker factor y from the speaker model with form of $s = \boldsymbol{m} + \boldsymbol{Vy} + \boldsymbol{Dz}$. Where m is denoted as UBM super-vector, V is an Eigen-voice rectangular matrix, D is a JFA parameter matrix and z is a JFA latent variable vector. JFA has two levels. First level models speakers under different conditions and second level uses first level to generate GMM. If all effective parameters on mean in second level are considered, GMM dependent expression would be as "Eq. (6)".

$$M_{ki} = m_k + U_k X_i + V_k y_{s(i)} + D_k Z_{ks(i)}, \tag{6}$$

where indices k shows various GMM components, I denotes as session, and $s(i)$ refers the speaker in session i. m_k, U_k, V_k, and D_k are the system parameters and X_i, $y_{s(i)}$, and $Z_{ks(i)}$ are hidden speaker and session variables. FA technique has computational overhead during estimating the hyper parameters of speaker-independent in training

phase and calculating model likelihoods in testing phase. While channel mismatch including any type of environmental mismatch (e.g., reverberation) is dependent, it is not efficient to use Eigen-channel compensation because it considers the independent channel distortion and decomposes the super-vector of test data to sum of training super-vector and channel distortion super-vector. While the mismatch is dependent between training and testing data, a new compensation technique called warp model is proposed in Ref. 110 for GMM-UBM. When super-vector of UBM is used instead of Eigen-channel compensation, this technique shows more improvement for channel compensation. However, in warp model compensation, the estimation of enrolment super-vector is sensitive for system reliability due to correlation estimation of distortion between enrolment and testing data.

Due to difficulty in training model for each speaker under all channels, SMS trains the parameters of model under channel variability by allowing to synthesis the speaker model for uncertain channel.

By applying SMS,[17] the GMM parameters of target would be adapted into a new channel condition. A channel-independent background model and channel-dependent adapted models must be transformed, if the new channel condition has not been considered in the training phase. SMS is model-domain and equivalent to FM. Moreover, SMS is required to extract mismatch information from test speech signal. However, in Ref. 111 channel mismatch compensation is proposed by learning a synthetic variance distribution (SVD) from stereo data which is recorded from multiple channels. PMC models noise by individual HMM model of the speech and noise, and it is similar to HMM decomposition. This technique has been applied for robust speaker verification by transforming Cepstral based features to log spectral domain before combining with mismatch function.[112] It seems effective for compensation under additive stationary and nonstationary noise. It might be unusual for GMM speaker identification due to unconstrained nature of the speech and the subsequent lack of temporal information which is captured by speaker model for modeling phoneme.

In contrast to JFA that apply high dimensional space to model super-vectors, a low dimensional super-vector is available (called i-vector[113]) that model within/between speaker variability. I-vector can model both speaker and channel variability as total variability space.

Although i-vectors are smaller to decrease complexity and cost, their performance is same as JFA technique. This is due to a large amount of data is applied to extract the i-vector. An i-vector can be represented as speaker and channel dependent GMM super-vector as "Eq. (7)"

$$\mu = m + Tw, \tag{7}$$

where m is similar UBM super-vector in JFA, T corresponding to a low-rank matrix. This matrix has shown major variability direction across whole data and w is random vector which has independent normal distribution. However, i-vector does not properly compensate the effects of channel variability as T is contains both speaker

and channel variability. That's why other SVM compensation approaches including WCCN and NAP have been used.

There are multiple compensation techniques using SVM such as Continuous Density SVM (CDSVM), Gaussian-mean Super-Vector (GSV), covariance kernel, Fisher kernel, Generalized linear discriminant sequence kernel (GLDSK), Polynomial dynamic time warping kernel (PDTWK), Term-Frequency Log-Likelihood Ratio (TFLLR), MLLR, and Pair-of-sequence (POS) found in literature.[114] The simplest methods GLDSK SVM[96] is computed as "Eq. (8)"

$$\mathrm{S} = \mathrm{W}_{target}^{T} b_{test}, \tag{8}$$

where $\mathbf{W}_{\mathbf{target}}^{\boldsymbol{T}}$ denotes the normalized model vector of the target speaker and $\boldsymbol{b}_{\mathbf{test}}$ denotes the normalized average expanded feature vector of the test utterance. Although, this method is efficient in computationally of verification phase, it is difficult to control the dimensionality of the super-vectors.[17]

The technique[29] combines SVM and GMM. The means of GMM are normalized by using variance to create high- and fixed-dimensional representation of an utterance with common coordinate system (super-vector), it is later fed to SVM training part. SVM classifies the features based on distant between PDFs as found in GMMs. GMM-UBM Mean Interval (GUMI) is another approach which uses PSK as kernel.[66] It uses Gaussian function instead of Gaussian means for producing super-vectors, and utilizes covariance matrices to extract the speaker's features. Other SVM-based approaches are probabilistic Distant kernels which use generative sequence models for SVM speaker modeling and Fisher kernels.[91] Linear method, as another SVM-based instance, transforms the input data to a new space by applying MLLR transform which is presented in "Eq. (9)".[17]

$$\mu_k' = \mathrm{A}\mu_k + \mathrm{b}, \tag{9}$$

where μ_k' is adapted mean vector, μ_k is the model mean vector, A and b are linear transform parameters. A and b modify EM algorithm by maximizing the likelihood of the training data.

Although, the GLDS-, GMM- and MLLR- SVM are usually and properly applied for modeling low-level features, the high level (e.g., prosodic and N-gram) use TFLLR[81] kernel for normalizing the original N-gram frequency by reversing square roots of the overall frequency of that N-gram. For noisy high-level features (phone tokenizer), soft binning technique is proposed in Ref. 78 by adding GMMs and Gaussian weights as the features for using in SVM super-vectors.

Two trends are available for normalization of super-vector in SVM which are dynamic range of features normalization and intersession variability compensation. Due to invariant property of SVM in linear transformations of feature space, variance normalization is applied to prevent domination of certain super-vector dimensions in inner product computation. Also, kernel-independent rank normalization technique is proposed in Ref. 115. This technique substitutes each feature by its relative position (rank) in the background data. The second trend intersession

(intra-speaker) variability compensation removes the unwanted variability from the super-vectors and enabling modeling channel condition for tackling condition where are not modeled in training data.

NAP[60,80] is a normalization technique based on SVM which uses the Eigen-channel. NAP has been applied WCCN which is used PCA to reweight each dimension. NAP uses P transformation matrix in feature space to eliminate the undesired variability, which is contaminated the GMMs super vector, in subspace features. The transformation matrix attenuates the nuisance features (e.g., channel variability) in the feature space by applying "Eq. (10)".

$$\mathrm{P} = \mathrm{I} - \mathrm{U}\mathrm{U}^{\mathrm{T}}, \tag{10}$$

where U is the matrix of Eigen-channel.

Applying multiple source, even each single one is not powerful, can properly improve the performance because more variability in training data enhance the robustness of speaker model.[10] In real situation, multiple distant microphones are combined to make use of information from different distant by using channel combination techniques.

There are four main combination techniques available for DSR systems including data combination (DC), Segment based Score Fusion (SSF), Frame based Score Competition (FSC), and Segment based Decision Voting (SDV).[10]

In DC technique, multiple channels data are combined to train the speaker models. In FSC, multiple GMMs are built over all channels (with total number C) for specific speaker k as $\{\Theta^k = \Theta^{k,ch1}, \ldots, \Theta^{k,chC}\}$. Then for every frame of test utterance in channel h, the GMM is compared with training GMMs of all channels except channel h as $\{\Theta^{k,ch1}, \ldots, \Theta^{k,chh-1}, \ldots, \Theta^{k,chh+1}, \ldots, \Theta^{k,chC}\}$.

The highest log likelihood score of all GMM models is selected as frame score. Finally, the log likelihood score of the entire test feature vector set from channel h is calculated by "Eq. (11)".

$$LL(X|\Theta^k) = \sum_{n=1}^{N} Max\{LL(x_n|\Theta^{k,chj})\}_{j=1,j\neq h}^{C}. \tag{11}$$

In contrast of FSC, the SSF technique applies whole test utterances (the entire set of test feature vectors X) and fusion weights (W_j) as "Eq. (12)" which they set to be equal across channel.

$$LL(X|\Theta^k) = \sum_{j=1,j\neq h}^{C} W_j \cdot LL(X|\Theta^{k,chj}). \tag{12}$$

Finally in SDV technique, specific speaker model is trained under one mismatched channel to model all feature vectors multiple times. The highest log likelihood score is selected from (C-1) number of the identity mismatched channel value.

8. Recognition (Scoring)

The recognition module applies the likelihood scores in determining whether the utterance belongs to target speakers or an imposter. Given X as a speech segment and S as a claimed identity, the speaker recognition system chooses one of the following hypotheses.

Hs: X is pronounced by S.
$H_{\bar{s}}$: X is not pronounced by S.

Then, likelihood scores are calculated by "Eq. (13)" where $p(X|H_s)$ and $p(X|H_{\bar{s}})$ are the PDF (integrated likelihood scores) which is estimated by the classifier and Q is a predefined value (known as threshold) for accepting or rejecting H_s.

$$\Lambda(X) = \frac{p(X|H_s)}{p(X|H_{\bar{s}})} \begin{cases} > \Theta \text{ accept } H_s \\ < \Theta \text{ accept } H_{\bar{s}}. \end{cases} \tag{13}$$

Determining the threshold Θ directly depends on speaker recognition application. It may be chosen based on environmental situation (noise), or level of demanding security. The threshold can reveal inter-speaker variability if it is more accurate. The developer must be aware of mismatch between test and development data to meet optimal point for a pre-determined threshold.

Typically, two thresholds should be considered for speaker verification and identification purpose. The first one separates the claimers and impostors by a predefined distant value and second one is determined level of matching between test speakers and claimer speakers.

Currently, different techniques have been proposed to score the i-vectors which is state-of-the-art technique in speaker recognition.[70] In cosine distant scoring, the angle between a test i-vector and a target i-vector is measured as "Eq. (14)":

$$S(X_{target}, X_{test}) = \frac{\langle X_{target}, X_{test} \rangle}{\|X_{target}\| \|X_{test}\|}. \tag{14}$$

In PLDA techniques such as (GPLDA and HTPLDA), the score between a test i-vector and a target i-vector is computed based on batch likelihood ratio as in "Eq. (15)":

$$S(X_{target}, X_{test}) = Ln \frac{P(X_{target}, X_{test}|H_1)}{P(X_{target}|H_0)P(X_{test}|H_0)}, \tag{15}$$

where $H1$ corresponds to while the speakers are same and, $H0$ denotes as the speaker are different.

9. Score Normalization

Multiple channel and environment effects (i.e., using various types of microphones for training and testing) might degrade the system performance. In contrast to

feature normalization which reduces linear channel effects, distant speech recording may induce other nonlinear degradations over speech signal. Not only score normalization techniques eliminates these types of distortions from GMM log-likelihood scores, but also compensates speaker-dependent score offset which is not compensated by model compensation and feature compensation techniques.

The main goal of score normalization is rescuing the score variability through channel under various situations. Score normalization is used in speaker recognition as "Eq. (16)":

$$\tilde{L}_\lambda(X) = \frac{L_\lambda(X) - \mu_\lambda}{\sigma_\lambda}, \tag{16}$$

where $L_\lambda(X)$ is the score for speech signal X and speaker model λ; μ_λ, σ_λ are the normalization parameters for speaker λ which are estimated from a (pseudo) impostor score distribution; and $\tilde{L}_\lambda(X)$ normalized score. Impostors' scores are assumed to have Gaussian distribution, with the mean and the standard deviation which depend on speaker model of the speech utterance. In the following different normalization techniques including Znorm,[77] Hnorm,[65] Tnorm,[116] and Dnorm[117] are discussed. Tnorm as well as Znorm try to minimize transmission channel and microphone effects.

Znorm. It uses imposter's speech similarity distribution to define normalized mean and variance parameters, when the test model established. It can be done during the training phase.

Tnorm. The mean and the standard deviation parameters are determined from the test utterance speech to enhance the performance, especially under low false alarm situation. The Tnorm evaluation part is generally online. In the Znorm normalization, Handset (Hnorm) and channel (Cnorm) of imposter parameters are used to test whole speaker model against channel-dependent set of imposters or a handset. In testing, the type of related handset or channel to the test utterance is detected. Then, the score normalization uses the corresponding sets of parameters.

HTnorm. As a generalization of TNorm, tests each test utterance against handset-dependent imposter models to estimate handset-dependent normalization parameters.

Dnorm. Despite Tnorm and Znorm, in Dnorm technique, pseudo imposter data is generated by using Monte Carlo method. Dnorm is applied where the imposter data is not available.

Although, speaker verification/identification applies score normalization techniques, speaker identification also can apply other techniques such as frame level normalization[118] and segmentation-and-normalization[119] approaches when large testing data from different environment are available.

Frame level normalization technique ranks speaker model likelihoods for each frame, then arbitrary weight function assigns weights to models. Instead of frame

likelihood value, accumulated weight scores is performed for classification. Another frame level approach is proposed in Ref. 120 to eliminate channel variability by a nonlinear transforming of frame score for each speaker. This approach can provide consistency between scores over all frames for the same speaker.

In segmentation-and-normalization technique, multiple GMM model training, which is based on various conditions, is applied to find the best likelihood (score) over all models and for each speaker. Then, normalization is done for allowing segment comparison to reduce the effect of mismatched condition.

Stable Score. Due to noise and reverberation in distant recording, the cumulative score is affected. The solution is selecting frames which are less affected by these factors. The main idea of stable scoring is the fact that the variation of acoustic features of a genuine speaker is lesser, when compared to a background speaker even in degradations. For this reason, a threshold over frame-wise of Euclidean in background speaker is applied to select frame. Then, the score is computed by the sum of scores of frames assigned to the claimant over the frames assigned to the claimant where the low score indicate the better chance for acceptance. Combination of this score normalization method with other type can improve the performance in distant speech recording case.[14]

10. DSR in Binaural Humanoids Robots

The redundancy of sensing signal can improve the robustness and performance of DSR systems.[121] Studies show that the multi-channel microphones improve EER compare to single channel from 29.5% to 12.7%.[122] Furthermore, it has demonstrated that the multi-channel microphone array system works better for a single localized noise than a diffuse noise field. This is result of gain which is nonzero over all undesirable directions. However, it is null for single noise source.[122] Basically, humanoid robots just embeds two microphones in artificial ears which are made them just like human. Although restricting the robot's acoustic sensor to a single pair of microphones is increased computational complexity, the main reasons of using binaural approach for robots are including embeddability, real-time, HRI, and investigation of human perception.[88] Studies show that simple binaural system still outperforms at least for 10% more than monaural system.[56] However, many issues might be consider for developing DSR for binaural robots e.g., speaker position which is presented based on two parameters such as azimuth in the horizontal plane and the elevation in the vertical place. Furthermore, combining of two noisy signals to a single less corrupted signal appropriately is another concern for binaural DSR researchers. In Ref. 123, mixture weight is optimized by cross-domain MFCCs combination of dual-microphone systems to improve the performance. Other methods are developed for binaural humanoid robot based on intercorrelation[56,59] (which two signals are correlated to obtain a signal) and concatenation (which is based on combination of left and right feature vectors). Although intercorrelation is

less sensitive to directional cues, concatenation seems to conserve more the directional information induced by Interaural Level Difference (ILD) or spectral notches.[59] It seems intercorrelation works better than concatenation method.[56,59] Also, some technique uses higher SNR (or called better ear) to select feature vector.[58] In Ref. 75, noise reduction and better-ear are combined together to improve DSR performance.

Computational auditory scene analysis (CASA) is an ability of humanoid robots to recognize and select a single target speaker who is following a conversation.[124] Although CASA is degraded under distant speech recording, a new concept, known as Missing Data (MD) classification is circumvented this degradation issue by classification of missing and unreliable acoustic information.[85] Binary mask is used to identify whether time–frequency unit is reliable or unreliable. The major binaural scene analyzer is ideal binary mask (IBM) which is required *a priori* knowledge about the azimuth location of the target sources. This knowledge makes a strong limitation for developing practical applications. In Refs. 58, 125 and 126, some MD techniques have been proposed to solve this problem.

11. Metrics for DSR Evaluation

Multiple metrics are utilized for performance evaluation of speaker recognition systems. Since the output of a speaker verification system is either acceptance or rejection of the authority of a claimer speaker, if a system reports an authorized speaker as an unauthorized speaker, it is counted as false rejection; however, if the system reports an unauthorized speaker as authorized speaker, it is counted as a false acceptance. The accuracy of a speaker verification system can be measured in terms of error ratios including FAR and FRR. EER is used for speaker recognition by finding a point where FAR equals to FRR. Even EER being smaller, it is better for recognition system. Notice that the EER increases for distant speech signals showing a decrease in the performance.

The value of decision threshold influences the security and user-friendly of system directly, and FAR and FRR indirectly. If a large threshold is chosen, the system accepts a few trusted speakers. It will increase the robustness and security of system, and FRR, as well. The opposite scenario may happen when a low threshold is chosen. In this case, user-friendly and FAR system increases and FRR decreases. A DET curve[127] assists in choosing the best threshold in the speaker recognition systems. The other performance metric is DCF (which is a weighted sum of FAR and FRR), minimum DCF (which is used for verifying the speaker by minimizing the value of DCF over the cross-validation set, while the decision threshold is changed), and actual DCF.

12. Results and Discussion

Previous studies had compared various types of speaker recognition systems.[16–19,57] However, none of them have deeper view on DSR systems. In this section, new results

Table 1. Average EER (%) for different VAD techniques.

System	Method	EER (%)
MFCC+SVM-GMM+i-vector[95]	AE-VAD	10.3
	ASR-VAD	8.88
	MR-VAD	9.61
	SM-VAD	6.28
	GMM-VAD	5.03
	SS+SM-VAD	4.8
	SS+AE-VAD	4.55
MFCC+GMM-UBM[128]	Energy	16.63
	LTSD	35.82
	Periodicity	16.76
MFCC+GMM-UBM[53]	G.729B	15.5233
	SMVAD	15.3020
	HNPNR	16.1300
	MEBTS	19.5040
	AEBTS	19.1453
	UBGME	15.7787

are presented to support theoretical information provided in earlier sections. These results have been obtained from rich published material.[3,19,45,54,55] Table 1 presents the average performance of each VAD techniques for different DSR systems. For first system, the results show that SS+AE-VAD is worked better than other VAD techniques, if its parameters are chosen properly. For second system, energy-based VAD is slightly better than periodicity. This is due to periodicity detection is seriously degraded under serious noise condition. For last system, although average performance for statistical-based VAD (SMVAD) is the best, bi-Gaussian log energy based VAD (UBGME) is outperformed than other VAD techniques for wide range of SNR. From the Table 1, it can be inferred that applying SS is very useful for energy-based VAD especially under low SNR condition. Furthermore, selecting reliable threshold strategy and removing noise as well as reverberation can significantly improve DSR performance.

A study is conducted on performance of various windowing techniques such as hamming, Thomson, Multipeak, and Sine Weighted Cepstrum Estimated (SWCE) in clean, 20, 10, 0, and −10 dB with average EERs of 11.658%, 10.176%, 10.288%, and 10.172% respectively. Based on these results, it seems SWCE slightly are more robust and simpler than Thomson on the noisier conditions.[100]

In Ref. 129, spectrum of MFCC is calculated by six different techniques such as FFT, LP, WLP, SWLP, XLP and SXLP for different SNRs of original, 20, 10, 0, and −10 dB. The average EERs 14.914%, 14.68%, 14.31%, 13.91%, 14.056%, and 14.116% are respectively. It can be concluded that the two stabilized versions such as SWLP or SXLP has best overall robustness under noisy condition and channel variation. However, SWLP technique performs well under severe noisy conditions. Table 2 shows the average EER for different spectrum techniques under various conditions.[129–131] As seen, white noise more seriously the performance of DSR

Table 2. Average EER (%) for different spectrum techniques under various conditions.

	With RASTA+CMVN				Without RASTA+CMVN			
	FFT	LP	WLP	SWLP	FFT	LP	WLP	SWLP
White noise	13.254	12.788	13.064	12.736	19.056	18.856	18.636	18.64
Factory noise	11.586	11.31	11.484	11.058	14.502	14.208	13.788	13.548
	With speech enhancement				Without speech enhancement			
	FFT	W-FFT	W-LP	W-MVDR	FFT	W-FFT	W-LP	W-MVDR
HVAC*	32.9	32.5	30.7	29.3	28	26.1	25.3	24.3
Crowd Noise	21.1	20.2	18.1	17.3	19.7	18.4	17.2	15.3

*Heating, ventilation and air-conditioning.

degrade than factory noise. Furthermore, WLP and SWLP work better than LP and FFT. Although WLP is better than SWLP for speech recognition, it seems SWLP works outperform than WLP for speaker recognition purpose.[45] From second part of Table 2, it can be inferred that warping spectral may improve the DSR performance. Also, MVDR spectrum seems more robust than other spectral techniques.

Table 2 also shows how applying feature normalization techniques can improve the performance in presence of noise. However, in case of W-MVDR speech enhancement cannot always improve the performance. As discussed earlier, the speaker-specific information may be removed by speech enhancement technique.

Table 3 compares different feature normalization techniques in terms of performance and time.[101,132] The first part of the Table 3 shows that both the STMVN and STMSN methods provide comparable speaker verification results using i-vectors to that of STG. Furthermore, STG is considerable more complex and takes longer time to normalize MFCC feature vectors compared to STMSN and STMVN. The second part of Table 3 concludes that feature mapping is more robust than other techniques. In addition, applying feature warping and feature mapping slightly outperforms in term of min DCF than feature mapping alone, but the feature mapping is no longer necessary to get good performance.

Table 4 shows performance of different score normalization techniques. As seen, ZT-norm outperforms than other score normalization techniques under noisy conditions.

Table 3. Averages EER (%) for different feature normalization techniques.

MFCC+i-vector+PLDA[101]	Method	EER (%)	Execution time (s)
	STG	3.06	16.05
	STMVN	3.17	3.01
	STMSN	3.4	1.71
PLP+GMM-UBM[132]	Method	EER (%)	Minimal DCF
	CMS	14.0338	4.7608
	Feature mapping	7.5169	3.1046
	Feature warping+Feature mapping	7.5938	3.0792

Table 4. Average EER (%) for different score normalization techniques.

System	Nonorm	Tnorm	Znorm	ZTnorm	TZnorm
MFCC+GMM+HMM[133]	2.19	1.44	1.65	1.505	1.775
MFCC+GMM-UBM[134]	10.8	12.0	9.4	8.6	9.4
PLP+GMM-UBM[132]	22.3167	15.7833	13.0833	11.7367	—

In Ref. 45, new investigation on DSR was performed to study changing minimum cost function as a function of threshold as seen in Fig. 4. This study uses database consisted of 45 (speaker 30 males and 15 females) with 4 omnidirectional microphones at distant of 0.06–0.13, 2, 4 and 6 ft in noisy lab environment with dimension of $20 \times 15 \times 10$ ft. It can be inferred that the minimum value of the cost function increases as the distant increases. Moreover, the cost function would be similar for all distance, if the variation of the feature is less. Increase in the feature variation leads to decrease in performance because of decrease in the confidence scores. The minimum cost for each distant is not the actual cost of the system at that distant since the threshold chosen for the system corresponds to a minimum cost of the close-speaking speech signals. Moreover, distant is not a parameter used during the verification phase. With the available features, additional information of distant will increase the performance of the system but finding distant from the speech signal is a difficult.

Table 5 compares classical speaker verification technique with respect to distant. It can be inferred that MFCC features shows more variation and degradation than short segment cepstral coefficients (SSCC) with increasing the distances. Moreover, strength of spectral peaks seem to be working better than energy for begin-end

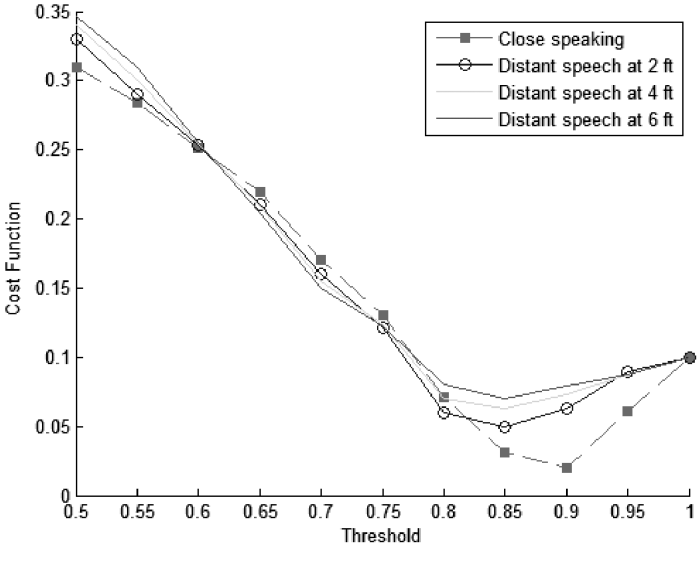

Fig. 4. Cost function versus threshold for different distant.[45]

Table 5. DSR system over different microphone to source distance.

System	Channel	EER
DTW+SSCC+high SRR[14]	< 1 ft,	1.43
	2 ft,	4.93
	4 ft,	9.84
	6 ft	9.93
DTW+MFCC+CMS[14]	< 1 ft,	2.32
	2 ft,	7.51
	4 ft,	11.01
	6 ft	12.03

detection of distant speech signal. Moreover, feature variation in genuine speaker is lesser than impostor speakers which is used by score normalization to improve the performance. However, performance significantly decreases, when distant increases. Furthermore, selecting frames with high signal to reverberation ratio (SRR) causes improving the performance. It must be noted that EERs are extracted directly from Ref. 45. More details are omitted for sakes of brevity.

Table 6 shows recognition rate with respect to distant by using different speaker modeling techniques with 25 speakers.[54] Three modeling are applied as least squared (LS), adaptive distribution (AD), and generalized gaussian model (GGM). It can be seen that the GGM recognition rate shows outperform than other modeling technique in close distant. But at distant AD seems to be more effective. Moreover, even by using exactly same speaker modeling technique, the performance is degraded with increasing the speaker to microphones distant due to ambient noise, reverberation, echo, head position and etc.

In Ref. 10 various DC methods have been studied by using 3D DMD, 2D DMD and ICSI database. Table 7 has been summarized averages for different DC. As seen, FSC is more effective than other techniques which may be due to combination level. In contrast of the other approaches which combine at the segment level, FSC combines multiple sources at the finest granularity, i.e., at the frame level.

Table 6. DSR system with different speaker modeling techniques at different distant.

System	Channel	Recognition rate (%)
MFCC+GGM	0.03 ft	93.85
MFCC+LS		93.45
MFCC+AD		93.25
MFCC+GGM	0.1 ft	64.45
MFCC+LS		73.9
MFCC+AD		74.6
MFCC+GGM	0.16 ft	64.4
MFCC+LS		74.6
MFCC+AD		77.8

Table 7. Different DC techniques.

Data combination techniques	Average recognition rate (%)
DC	54.8
FSC	60.06
SSF	49.03
SDV	42.03

Table 8. Distant speaker identification methods.

System	Recognition rate (%)
Single-channel GMM[10]	73.4
Multi-channel GMM[10]	94.7
High level (Phonetic)[10]	86.7
High level (syllabic) GMM[12]	97.14
High level (syllabic) HMM[12]	96.91
High level (syllabic) GMM+HMM[12]	99.04
MFCC+UBM+GMM[135]	88.64
MFCC+SVM[135]	77.08
MFCC+RVM[135]	80.71

Table 8 compares the performances of different DSR systems. As seen in Table 8, the phonetic method is outperformed than single-channel GMM but multi-channel GMM gives better results as compared to phonetic. Moreover, the fusion between GMM and the phonetic systems did not give any additional gain.[10] Also, applying high-level linguistic feature (e.g., syllables) has given the best performance as compared to other systems due to capture speaker idiosyncrasies by using n-gram. Although high-level features are more robust for DSR systems especially under the mismatch condition, large training data must be fed to the systems to reliably estimate phonetic n-gram models.

It can be shown that the generative model is more applicable/productive for speaker identification than discriminative model. As seen in Table 4, GMM-UBM outperforms than relevance vector machine (RVM) and SVM. However, RVM performed better than SVM. By fusing, the performance is improved.

Figure 5 shows the speaker identification performance under various amount of factory noise. As seen, the performance of MFCC mono is decreased with decreasing SNR. MFCC binaural is working better than mono due to using better SNR for recognition. The performance of MFCC binaural can be improved by applying noise reduction technique.[58]

Table 9 presents performance of DSR system in term of EER for various scoring techniques. As seen, the fastest scoring technique is linear scoring method with approximately reasonable EER.[61] Also, it shows PSVM slightly outperform than GPLDA with similar time.[89]

In Ref. 136, PLDA and cosine distant scoring are compared which is shown that the average EER for PLDA (3.61%) is significantly less than cosine distance scoring

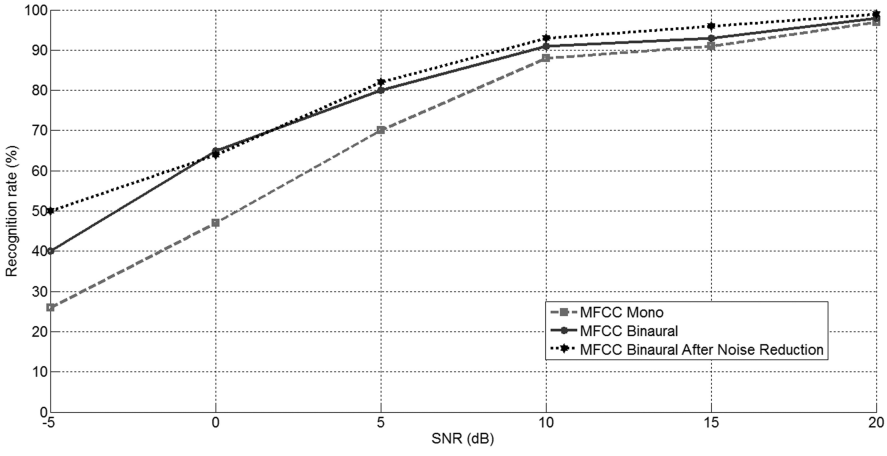

Fig. 5. Average speaker recognition rate (%) for mono, binaural, and noise reduction in presence of different interfering factory noise.

Table 9. Average EER (%), time, and RTF for different scoring techniques.

Method	EER (%)	Time (s)	Real-time factor (RTF)
Frame-by-frame	3.27	1010	1.6e-1
Integration	2.21	50	7.93e-3
Point estimate	3.54	160	2.54e-2
Point estimate LPT	9.61	36	5.71e-3
Linear scoring	3.55	13	2.07e-3
PLDA	5.21	N/A	N/A
Gender-dependent PLDA	4.68	N/A	N/A
Heavy Tailed PLDA	4.85	N/A	N/A
Pairwise SVM	4.54	N/A	N/A

(5.09%) techniques for all i-vector extraction techniques. This strength of PLDA is due to the uncompensated i-vector behavior which is heavy tailed, and this heavy-tailed can explicitly model outliers in the i-vector space. Recently, a new approach, known as length-normalized Gaussian PLDA (GPLDA), is introduced which is more efficient computationally than PLDA.[70]

Table 10 compares quality of the different binary mask techniques which shows that the performance for SNR binary mask[126] is near to IBM[137] in both anechoic and reverberant situations.

Figure 6 shows the recognition rates (%) for binaural and monaural DSR systems. As seen, whenever the length of the signal is increased, the performance of DSR systems are increased. Furthermore, SNR seems to have significant contribution to DSR performance. While the SNR is decreased, the performance is decreased for both binaural and monaural DSR systems. Finally, it can be inferred that the binaural DSR system is outperformed than monaural due to the use of two separate signals which may works better for signals with various SNRs.

Table 10. Recognition rate (%) for different binary mask techniques.

Method	Anechoic	Reverberation (RT = 0.29 s)
Without *a priori* knowledge[58]	92.4	85.5
With *a priori* knowledge[138]	95.7	88.5
SNR binary mask[126]	96.4	92.5
ITD Palomäki[125]	90.2	77.3
ITD & ILD Palomäki[125]	79.4	61.4
Ideal binary mask[137]	97.8	96.8

Note: ILD (Interaural Level Difference); IPD/ITD (Interaural Phase Difference/Interaural Time Difference).

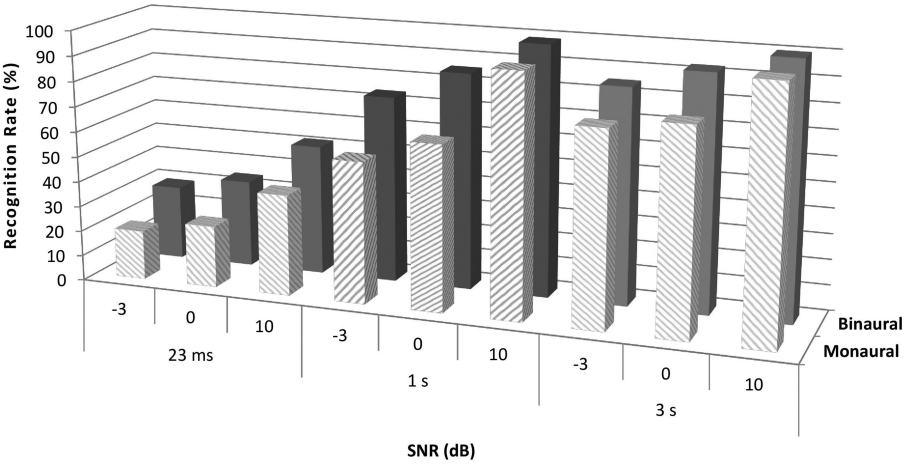

Fig. 6. Recognition rate (%) for binaural and monaural DSR system under different SNRs and lengths.

Table 11 shows recognition rate for different training and testing directions in terms of azimuth in the horizontal plane and the elevation in the vertical place which are new parameters in binaural signal.[26,59] From the Table 11, it can be concluded that the best recognition rate is obtained only in the direction which is already trained.

For better investigation on speaker position, training system, length and SNR of the binaural signal, Fig. 7 presents recognition rate for these parameters. As seen, the best recognition rates are achieved when multiple directions are trained to DSR

Table 11. The recognition rate (%) for different training and testing directions.

Training/Testing	(0,0)	(−45,45)	(45,90)
(0,0)	47	35	45
(−45,45)	35	42	32
(45,90)	45	30	50

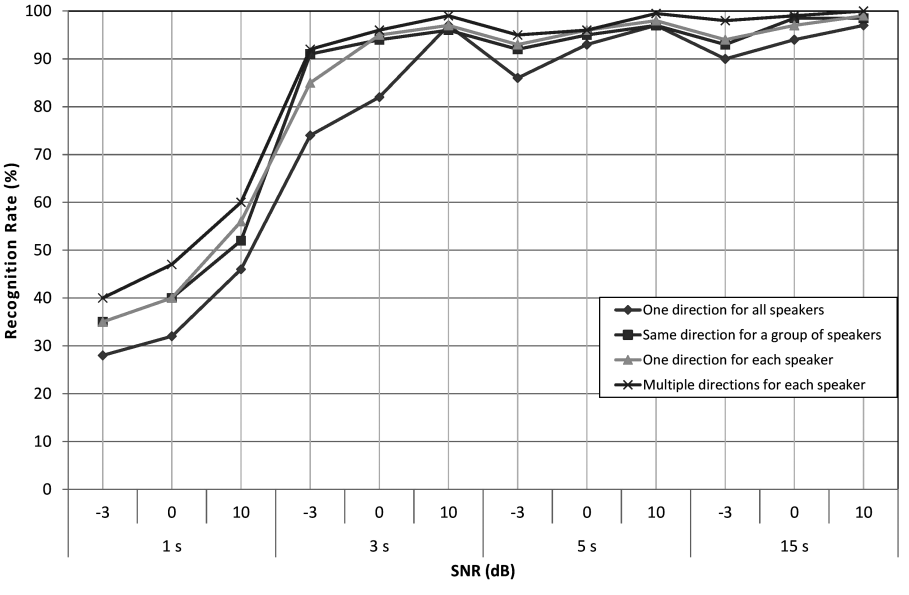

Fig. 7. Recognition rate (%) for different training and testing scenarios under different SNRs and different lengths.

systems. Also, it seems whenever the diversity for speaker positions are increased in training system, the performance is improved. It is seen that like speaker feature, speaker position can also be trained to DSR system. Furthermore, whenever the signal SNR and length is increased, the DSR performance is increased.

13. Dereverberation and DSR Databases

Many researchers have reported tremendous degradation on DSR performance under reverberation when the results are compared with baseline systems (i.e., the EER is 6.79% for clean speech changed up to 18.32% for reverberant test[83]). Due to this, the requirement for constructing robust system is inevitable. Actually, reverberation is highly dependent on convolutive distortion in such way that clean speech is convolved with RIR. Microphone array can be used to reduce this effect, but it is not always feasible and costly.

Feature normalization, score normalization and model compensation only can suppress the reverberation with short-duration RIR (reverberation time (RT) is shorter than 0.5 s) like communication channel effect. However, they might not compensate the effect of long-duration RIR (RT is more than 1.5 s). In Ref. 139, ANN is performed dereverberation on single channel by mapping the reverberate mel-spectral feature to anechoic feature. Some researches try to overcome long-duration RIR by training the speaker model with reverberant data.[83] Matching RT between training and testing can improve the EER from 16.44% to 9.9% with

Znorm and Tnorm.[83] In Ref. 63, the author concludes that reverberant data may decrease the GMM speaker model order due to the Gaussians means getting closer together when RT increases. This is important because it assumed that speaker model order is constant. In highly reverberant environment, the EER will be reduced if the model order is reduced. Also, applying feature normalization like CMS for dereverberation can degrade the performance because it does not consider the length of the RIR.[140] In Ref. 83 reverberation is classified by applying reverberant background model (RBM) and frame-definition (FD). Also, it has shown that RT might not be the best parameter in all conditions.

Room has many parameters such as RT, early decay time (EDT), clarity index (C80), definition (D50), and center time (Ts) that can be derived directly from RIR. These parameters can be applied for dereverberation.

However, some independent parameters such as the volume of the room, absorption (or reflection), and source-to receiver distant affect the RIR. It must be noted that extraction of these parameters are difficult and require feature extraction.

Some researchers have been tried to classify the room-volume from reverberant speech[140,141] or RIR[142–145] by applying statistical pattern recognition method to improve DSR performance. Some robot system use acoustic model likelihood criterion to estimate room transfer function (RTF) for the signal dereverbration.[146] Although room-volume classification can improve the performance under reverberation condition, many parameters such as difference absorption, source to receiver distant are needed. A new method is proposed to improve the DSR performance by classification of room-volume without need of these parameters.[147] Another blind dereverberation technique is proposed which process the outputs of two microphones by applying Cepestra operation and signal reconstruction based on the phase. Also, filter bank is applied on reverberant speech to recover the envelope modulation. A new technique is proposed base on generalized singular value decomposition (GSVD) or generalized eigen value decomposition (GEVD) for constructing data matrixes in presence of color noise.[148] Some techniques[10] assumed that reverberation is treating same additive noise and they have been applied SS to suppress it. But they cannot estimate the optimum parameters for SS. However, time domain LPC is applied to attenuate the late reverberation in spectral domain by using SS. Finally in Ref. 12, multi-channel least mean square (MCLMS) dereverberation method is proposed that estimate the spectrum of impulse response for the SS in a frequency domain blindly. Different possibility including linear post-filtering and FSC are proposed for DSR.[15] Post filtering uses linear time invariant (LTI) filter on features for minimizing square distortion between training and testing sets. The main advantage of post filtering over Wiener filter is that post filter is invariant and unique on detailed nature of environment and reverberation. Also, four scenarios (1) match condition, (2) mismatch condition with multiple microphone, (3) mismatch condition with single microphone (4) mismatch condition with single changing microphone position for evaluation of FCS are investigated. It can be concluded that: first using the FSC dereverberation is better than baseline system; second proper selection of

Table 12. Databases for DSR systems.

Database name	Number of speaker	Description
International Computer Science Institute (ICSI)[149]	53	Location: Berkeley, CA 75 meetings with simultaneous multichannel audio recordings Table microphones: Desktop omnidirectional pressure zone microphones (PZMs) aligned in a staggered line along the table center.
2-D Distant Microphone database (2D DMD)[9]	30 (16 female, 14 male)	Location: Interactive Systems Labs Recordings from 8 microphones at various distances Microphones aligned in a staggered line along the table center.
3D Distant Microphone Database (3D DMD)[9]	24 (4 female, 20 male)	8 distant microphones (5 microphones hanging from the ceiling, 3 microphones mounted on the meeting table) Positioned in the 3D space.
Far-Field Speaker Identification (FarSID)[15]	10	Location: Carnegie Mellon University 16 microphones (5 hanging from the ceiling, 7 aligned in a microphone array (mounted on the table) with 0.16 ft distant from each other, 2 remained microphones mounted on the table only) Positioned in 3D space.
Head Related Transfer Function (HRTF)[150]	45	Location: University of California, CIPIC Interfaces Lab. Release 1.0 includes head-related impulse responses 25 different azimuths and 50 different elevations for 45 subjects (1250 directions) at approximately 5" angular increments.
Computers in the Human Interaction Loop (CHIL) seminar database[151]	> 26	Distant speech recordings of five different scientific seminars A NIST Mark III 64-channel microphone array mounted on one wall of the seminar rooms.
Multi-Channel Wall Street Journal Audio Visual Corpus (MC-WSJ-AV)[152]	The WSJCAM0 development set contains 20 speakers and the evaluation-1 and evaluation-2 sets 14 speakers each.	Strong reverberation, ambient noise and non-stationary noise Locations: The Centre for Speech Technology Research, Edinburgh (UEDIN), the IDIAP Research Institute, Switzerland (IDIAP) and TNO Human Factors, the Netherlands (TNO). data recorded at the UEDIN site with three conditions: 1. Single Speaker Stationary. 2. Single Speaker Moving. 3. Overlapping Speakers (Stationary).

microphone positions does not affect the FSC performance and third, FSC can be improved under multiple microphone.

Many speech databases are available for speaker recognition purpose (for more detail see Ref. 16), such as YOHO, TIMIT, SIVA, KING, SPIDER, POLYCOST, PolyVar, and AHUMADA.

However, other database must be employed for DSR or robotic application because other factors, such as microphone positioning, microphone distant to speaker, room characteristics, and speaking style are going to be evaluated. It must be noted that some databases are prepared by corrupting artificially the clean databases (e.g., the NTIMIT database that uses carbon button transducers and actual telephone line conditions to suffer TIMIT speech database). Table 12 presents major databases for DSR application.

14. Conclusions and Future Trends

This paper reviewed different DSR techniques phase by phase. As discussed, different degradation factor such as coloration, ambient noise and reverberation can decrease the DSR performance. For this reason, different state-of-the-art phases in a DSR system were explained in detail to improve the DSR performance. Furthermore, the results were presented for each DSR technique to support the theoretical background. Based on the results, it was shown that whenever the source to microphone distant is increased, the DSR performance is decreased. The real world DSR is a new challenge and opportunity for researchers who wish to develop robust DSR for real conditions which are different and unpredictable compared with lab conditions. DSR systems have multiple limitations and these limitations will motivate researchers to investigate more rigorously in this area. For further works, a researcher can study and design DSR systems by using high-level features which are more stable than low-level features against noise and masquerade, and are more accurate and robust. Finding the best domain (log-spectral or Cepstral domains) for modeling background noise and reverberation as well as robust feature, which suffer lesser variation with distance, can be a future trend in this field.

References

1. A. K. Jain, A. Ross and S. Prabhakar, An introduction to biometric recognition, *IEEE Trans. Circuits Syst. Video Technol.* **14**(1) (2004) 4–20.
2. J. Cao *et al.*, The visual-audio integrated recognition method for user authentication system of partner robots, *Int. J. Humanoid Robot.* **8**(04) (2011) 691–705.
3. R. Bolle, *Guide to Biometrics* (Springer, NY, 2004).
4. Q. Mao, X. Wang and Y. Zhan, Speech emotion recognition method based on improved decision tree and layered feature selection, *Int. J. Humanoid Robot.* **7**(02) (2010) 245–261.
5. C. Liu *et al.*, Generation of nodding, head tilting and gazing for human–robot speech interaction, *Int. J. Humanoid Robot.* **10**(01) (2013) 1350009.
6. J. Weng, Developmental robotics: Theory and experiments, *Int. J. Humanoid Robot.* **1** (02) (2004) 199–236.
7. H. G. Okuno *et al.*, Computational auditory scene analysis and its application to robot audition, *IEEE Int. Conf. Informatics Research for Development of Knowledge Society Infrastructure, 2004. ICKS 2004* (2004).
8. M. Wolfel and J. McDonough, *Distant Speech Recognition* (Wiley-Blackwell, US, 2009).
9. Q. Jin, *Robust Speaker Recognition*, PhD diss., Carnegie Mellon University (2007).

10. Q. Jin, T. Schultz and A. Waibel, Far-field speaker recognition, *IEEE Trans. Audio, Speech Lang. Process.* **15**(7) (2007) 2023–2032.
11. Q. Jin, R. Li, Q. Yang, K. Laskowski and T. Schultz, Speaker identification with distant microphone speech, *IEEE Trans. Conf. Acoustics Speech Signal Processing (ICASSP), 2010* (IEEE International Conference, 2010), pp. 4518–4521.
12. L. Wang, *A Study on Hands-Free Speech/Speaker Recognition*, PhD diss., PhD Thesis, Toyohashi University of Technology (2008).
13. L. Wang, N. Kitaoka and S. Nakagawa, Robust distant speaker recognition based on position-dependent CMN by combining speaker-specific GMM with speaker-adapted HMM, *Speech Commun.* **49**(6) (2007) 501–513.
14. B. Avinash, Exploring features for text-dependent speaker verification in distant speech signals, Master thesis, International Institute of Information Technology India (2006).
15. Q. Jin *et al.*, Compensation approaches for far-field speaker identification, in *NIST SRE Workshop* (2008).
16. A. Fazel and S. Chakrabartty, An overview of statistical pattern recognition techniques for speaker verification, *IEEE Circuits Syst. Mag.* **11**(2) (2011) 62–81.
17. T. Kinnunen and H. Li, An overview of text-independent speaker recognition: From features to supervectors, *Speech Commun.* **52**(1) (2010) 12–40.
18. R. Togneri and D. Pullella, An overview of speaker identification: Accuracy and robustness issues, *IEEE Circuits Syst. Mag.* **11**(2) (2011) 23–61.
19. D. A. Reynolds, An overview of automatic speaker recognition technology, *IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), 2002* (IEEE, 2002), pp. 4072–4075.
20. J. L. Flanagan, *Speech Analysis: Synthesis and Perception* (Springer-Verlage, NY, 1972).
21. B. S. Atal, Automatic recognition of speakers from their voices, *Proc. IEEE*, Vol. 64, No. 4 (1976), pp. 460–475.
22. L. Wang, N. Kitaoka and S. Nakagawa, Robust distant speech recognition by combining multiple microphone-array processing with position-dependent CMN, *EURASIP J. Appl. Signal Process.* **2006** (2006) 204–204 1–11.
23. M. Möser, *Engineering Acoustics: An Introduction to Noise Control* (Springer, NY, 2009).
24. U. H. Kim, J. Kim, D. Kim, H. Kim and B. J. You, Speaker localization using the TDOA-based feature matrix for a humanoid robot, *17th IEEE Int. Symp. Robot and Human Interactive Communication, 2008. RO-MAN 2008* (IEEE, 2008), pp. 610–615.
25. J. Sanchez-Riera, X. Alameda-Pineda, J. Wienke, A. Deleforge, S. Arias, J. Cech, S. Wrede and R. Horaud, Online multimodal speaker detection for humanoid robots, in *12th IEEE-RAS Int. Conf. Humanoid Robots (Humanoids), 2012* (IEEE, 2012), pp. 126–133.
26. B. Breteau, S. Argentieri, J.-L. Zarader, Z. Wang and K. Youssef, Binaural speaker recognition for humanoid robots, in *IEEE Int. Conf. Robotics and Biomimetics (ROBIO), 2010* (IEEE, 2010), pp. 1405–1410.
27. K. Youssef, S. Argentieri and J.-L. Zarader, From monaural to binaural speaker recognition for humanoid robots, in *10th IEEE-RAS Int. Conf. Humanoid Robots (Humanoids), 2010* (IEEE, 2010), pp. 580–586.
28. A. Ariyaeeinia *et al.*, A test of the effectiveness of speaker verification for differentiating between identical twins, *Sci. Justice* **48**(4) (2008) 182–186.
29. D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin *et al.*, The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition, in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, 2003 (ICASSP'03)*, Vol. 4 (IEEE, 2003), pp. IV–784.

30. J. P. Campbell, D. A. Reynolds and R. B. Dunn, Fusing high-and low-level features for speaker recognition, in *INTERSPEECH* (2003), pp. 2665–2668.
31. J. McLaughlin, D. A. Reynolds and T. P. Gleason, A study of computation speed-UPS of the GMM-UBM speaker recognition system, in *EUROSPEECH*, Vol. 99 (1999), pp. 1215–1218.
32. T. Kinnunen, E. Karpov and P. Franti, Real-time speaker identification and verification, *IEEE Trans. Audio, Speech Lang. Process.* **14**(1) (2006) 277–288.
33. G. Sarkar and G. Saha, Efficient pre-quantization techniques based on probability density for speaker recognition system, in *IEEE Region 10 Conf. TENCON 2009-2009* (IEEE, 2009), pp. 1–6.
34. M. Wölfel, A joint particle filter and multi-step linear prediction framework to provide enhanced speech features prior to automatic recognition, in *Hands-Free Speech Communication and Microphone Arrays, 2008 (HSCMA 2008)* (IEEE, 2008), pp. 132–135.
35. L. Ferrer, M. Graciarena, A. Zymnis and E. Shriberg, System combination using auxiliary information for speaker verification, in *IEEE Int. Conf. Acoustics, Speech and Signal Processing, 2008 (ICASSP 2008)* (IEEE, 2008), pp. 4853–4856.
36. B. Atal, Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification, *J. Acoustical Soc. Am.* **55**(6) (1974) 1304–1312.
37. D. O'shaughnessy, *Speech Communication: Human and Machine* (Universities Press, 1987).
38. D. Reynolds, Channel robust speaker verification via feature mapping, in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, 2003 (ICASSP'03)*, Vol. 2 (IEEE, 2003), pp. II–53.
39. H. Hermansky, Perceptual linear predictive (PLP) analysis of speech, *J. Acoustical Soc. Am.* **87** (1990) 1738.
40. M. D. Skowronski and J. G. Harris, Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition, *J. Acoustical Soc. Am.* **116** (2004) 1774.
41. Y. Nishimura, T. Shinozaki, K. Iwano and S. Furui, Noise — robust speech recognition using multiband spectral features, *The Journal of the Acoustical Society of America* **116**(4) (2004) 2480–2490.
42. K. Laskowski and Q. Jin, Modeling instantaneous intonation for speaker identification using the fundamental frequency variation spectrum, in *IEEE Int. Conf. Acoustics, Speech and Signal Processing, 2009 (ICASSP 2009)* (IEEE, 2009), pp. 4541–4544.
43. Z. Yu, S. Hongbin, W. Junjie and Y. Yonghong, An improved speaker diarization system for multiple distance microphone meetings, in *Fifth Int. Conf. Intelligent Computation Technology and Automation (ICICTA), 2012* (IEEE, 2012), pp. 80–83.
44. S. Dharanipragada, U. H. Yapanel and B. D. Rao, Robust feature extraction for continuous speech recognition using the MVDR spectrum estimation method, *IEEE Trans. Audio, Speech Lang Process.* **15**(1) (2007) 224–234.
45. R. Saeidi *et al.*, Temporally weighted linear prediction features for tackling additive noise in speaker verification, *IEEE Signal Process. Lett.* **17**(6) (2010) 599–602.
46. C. Magi *et al.*, Stabilised weighted linear prediction, *Speech Commun.* **51**(5) (2009) 401–411.
47. C. Hanilçi *et al.*, Regularized all-pole models for speaker verification under noisy environments, *IEEE Signal Process. Lett.* **19**(3) (2012) 163–166.
48. T. Kinnunen and P. Alku, On separating glottal source and vocal tract information in telephony speaker verification, in *IEEE Int. Conf. Acoustics, Speech and Signal Processing, 2009 (ICASSP 2009)* (IEEE, 2009), pp. 4545–4548.

49. N. Zheng, T. Lee and P. Ching, Integration of complementary acoustic features for speaker recognition, *IEEE Signal Process. Lett.* **14**(3) (2007) 181–184.
50. K. S. R. Murty and B. Yegnanarayana, Combining evidence from residual phase and MFCC features for speaker recognition, *IEEE Signal Process. Lett.* **13**(1) (2006) 52–55.
51. S. Mahadeva Prasanna, C. S. Gupta and B. Yegnanarayana, Extraction of speaker-specific excitation information from linear prediction residual of speech, *Speech Commun.* **48**(10) (2006) 1243–1261.
52. B. Hanson and T. Applebaum, Robust speaker-independent word recognition using static, dynamic and acceleration features: Experiments with Lombard and noisy speech, in *Int. Conf. Acoustics, Speech, and Signal Processing, 1990 (ICASSP-90)* (IEEE, 1990), pp. 857–860.
53. M. Sahidullah and G. Saha, Comparison of speech activity detection techniques for speaker recognition, arXiv:1210.0297.
54. L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition* (Prenctice Hall: Englewood Cliffs, NJ, 1993).
55. X. Lu and J. Dang, An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification, *Speech Commun.* **50**(4) (2008) 312–322.
56. K. Youssef, B. Breteau, S. Argentieri, J.-L. Zarader and Z. Wang, Approaches for automatic speaker recognition in a binaural humanoid context, in *ESANN* (2011), pp. 1–6.
57. C. H. Lee, F. K. Soong and K. K. Paliwal, *Automatic Speech and Speaker Recognition: Advanced Topics*, Vol. 355 (Springer, NY, 1996).
58. T. May, S. van de Par and A. Kohlrausch, A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation, *IEEE Trans. Audio, Speech Lang. Process.* **20**(7) (2012) 2016–2030.
59. K. Youssef, S. Argentieri and J.-L. Zarader, Binaural speaker recognition for humanoid robots, in *2010 11th Int. Conf. Control Automation Robotics & Vision (ICARCV)* (IEEE, 2010), pp. 2295–2300.
60. Q. Jin, J. Navratil, D. Reynolds, J. P. Campbell, W. D. Andrews and J. S. Abramson, Combining cross-stream and time dimensions in phonetic speaker recognition, in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, 2003 (ICASSP'03)*, Vol. 4 (IEEE, 2003), pp. IV–800.
61. O. Glembek, L. Burget, N. Dehak, N. Brümmer and P. Kenny, Comparison of scoring methods used in speaker recognition with joint factor analysis, in *IEEE Int. Conf. Acoustics, Speech and Signal Processing, 2009 (ICASSP 2009)* (IEEE, 2009), pp. 4057–4060.
62. E. Shriberg, L. Ferrer, S. S. Kajarekar, N. Scheffer, A. Stolcke and M. Akbacak, Detecting nonnative speech using speaker recognition approaches, in *Odyssey* (2008), p. 26.
63. L. Wang, Y. Kishi and A. Kai, Distant speaker recognition based on the automatic selection of reverberant environments using GMMs, in *Chinese Conf. Pattern Recognition, 2009 (CCPR 2009)* (IEEE, 2009), pp. 1–5.
64. N. R. Shabtai, Y. Zigel and B. Rafaely, The effect of GMM order and CMS on speaker recognition with reverberant speech, in *Hands-Free Speech Communication and Microphone Arrays, 2008 (HSCMA 2008)* (IEEE, 2008), pp. 144–147.
65. D. Reynolds, The effects of handset variability on speaker recognition performance: Experiments on the switchboard corpus, in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing, 1996 (ICASSP-96)*, Vol. 1 (IEEE, 1996), pp. 113–116.

66. W. D. Andrews, M. A. Kohler, J. P. Campbell, J. J. Godfrey and J. Hernández-Cordero, Gender-dependent phonetic refraction for speaker recognition, in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), 2002*, Vol. 1 (IEEE, 2002), pp. I–149.
67. R. E. Slyh, E. G. Hansen and T. R. Anderson, Glottal modeling and closed-phase analysis for speaker recognition, in *ODYSSEY04-The Speaker and Language Recognition Workshop* (2004), pp. 315–322.
68. A. O. Hatch, B. Peskin and A. Stolcke, Improved phonetic speaker recognition using lattice decoding, in *ICASSP (1)* (2005), pp. 169–172.
69. T. Yamada, A. Tawari and M. M. Trivedi, In-vehicle speaker recognition using independent vector analysis, in *15th Int. IEEE Conf. Intelligent Transportation Systems (ITSC), 2012* (IEEE, 2012), pp. 1753–1758.
70. A. Kanagasundaram *et al.*, I-vector based speaker recognition using advanced channel compensation techniques, *Comput. Speech Lang.* **28**(1) (2014) 121–140.
71. G. S. Kajarekar, E. Shriberg, K. Sönmez, A. Stolcke and A. Venkataraman, Modeling duration patterns for speaker recognition (2003).
72. A. G. Adami, R. Mihaescu, D. Reynolds and J. J. Godfrey, Modeling prosodic dynamics for speaker recognition, in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, 2003 (ICASSP'03)*, Vol. 4 (IEEE, 2003), pp. IV–788.
73. E. Shriberg *et al.*, Modeling prosodic feature sequences for speaker recognition, *Speech Commun.* **46**(3) (2005) 455–472.
74. C. Vielhauer *et al.*, Multimodal speaker authentication–evaluation of recognition performance of watermarked references, in *Proc. 2nd Workshop on Multimodal User Authentication (MMUA)* (Toulouse, France, 2006).
75. T. May, S. van de Par and A. Kohlrausch, Noise-robust speaker recognition combining missing data techniques and universal background modeling, *IEEE Trans. Audio, Speech Lang. Process.* **20**(1) (2012) 108–121.
76. Y. Konig, L. Heck, M. Weintraub and K. Sonmez, Nonlinear discriminant feature extraction for robust text-independent speaker recognition, in *Proc. RLA2C, ESCA workshop on Speaker Recognition and its Commercial and Forensic Applications* (1998), pp. 72–75.
77. K.-P. Li and J. E. Porter, Normalizations and selection of speech segments for speaker recognition scoring, in *Int. Conf. Acoustics, Speech, and Signal Processing, 1988 (ICASSP-88)* (IEEE, 1988), pp. 595–598.
78. L. Ferrer, E. Shriberg, S. Kajarekar and K. Sonrnez, Parameterization of prosodic feature distributions for SVM modeling in speaker recognition, in *IEEE Int. Conf. Acoustics, Speech and Signal Processing, 2007 (ICASSP 2007)*, Vol. 4 (IEEE, 2007), pp. IV–233.
79. M. A. Kohler, W. D. Andrews, J. P. Campbell and J. Herndndez-Cordero, Phonetic speaker recognition, in *Conference Record of the Thirty-Fifth Asilomar Conference on Signals, Systems and Computers, 2001*, Vol. 2 (IEEE, 2001), pp. 1557–1561.
80. J. Navrátil, Q. Jin, W. D. Andrews and J. P. Campbell, Phonetic speaker recognition using maximum-likelihood binary-decision tree models, in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, 2003 (ICASSP'03)*, Vol. 4 (IEEE, 2003), pp. IV–796.
81. W. M. Campbell *et al.*, Phonetic speaker recognition with support vector machines, in *Advances in Neural Information Processing Systems*, Vol. 16 (Cambridge, MA, 2003).
82. J. Navrátil, Q. Jin, W. D. Andrews and J. P. Campbell, Phonetic speaker recognition using maximum-likelihood binary-decision tree models, in *Proc. IEEE Int. Conf.*

Acoustics, Speech, and Signal Processing, 2003 (ICASSP'03), Vol. 4 (IEEE, 2003), pp. IV–796.
83. I. Peer, B. Rafaely and Y. Zigel, Reverberation matching for speaker recognition, in *IEEE Int. Conf. Acoustics, Speech and Signal Processing, 2008 (ICASSP 2008)* (IEEE, 2008), pp. 4829–4832.
84. L. Wang, N. Kitaoka and S. Nakagawa, Robust distant speech and speaker recognition based on position dependent cepstral mean normalization, *INTERSPEECH 2005 — Eurospeech, 9th European Conference on Speech Communication and Technology*, Lisbon, Portugal, September 4–8, 2005, pp. 1–4.
85. D. Pullella, M. Kuhne and R. Togneri, Robust speaker identification using combined feature selection and missing data recognition, in *IEEE Int. Conf. Acoustics, Speech and Signal Processing, 2008 (ICASSP 2008)* (IEEE, 2008), pp. 4833–4836.
86. Y. Shan and J. Liu, Robust Speaker Recognition in Cross-Channel Condition, in *2nd Int. Congress on Image and Signal Processing, 2009 (CISP'09)* (IEEE, 2009), pp. 1–5.
87. K. S. Rao, A. K. Vuppala, S. Chakrabarti and L. Dutta, Robust speaker recognition on mobile devices, in *Int. Conf. Signal Processing and Communications (SPCOM), 2010* (IEEE, 2010), pp. 1–5.
88. I. A. McCowan, J. Pelecanos and S. Sridharan, Robust speaker recognition using microphone arrays, in *2001: A Speaker Odyssey-The Speaker Recognition Workshop* (2001), pp. 1–6.
89. S. Cumani, Speaker and language recognition techniques, Ph.D. thesis, Politecnico di Torino (2012).
90. Q. Lin, E. Jan, C. Che and J. L. Flanagan, Speaker identification in teleconferencing environments using microphone arrays and neural networks, in *Automatic Speaker Recognition, Identification and Verification* (1994), pp. 235–238.
91. G. R. Doddington, Speaker recognition based on idiolectal differences between speakers, in *INTERSPEECH* (2001), pp. 2521–2524.
92. L. Maruti, R. B. Rao and V. Sagvekar, Article: Speaker Recognition using VQ and DTW, *IJCA Proceedings on International Conference on Advances in Communication and Computing Technologies 2012 ICACACT(3)* (August 2012), pp. 18–20.
93. R. P. Ramachandran *et al.*, Speaker recognition — general classifier approaches and data fusion methods, *Pattern Recognit.* **35**(12) (2002) 2801–2821.
94. A. Stolcke, E. Shriberg, L. Ferrer, S. Kajarekar, K. Sonmez and G. Tur, Speech recognition as feature extraction for speaker recognition, in *IEEE Workshop on Signal Processing Applications for Public Security and Forensics, 2007 (SAFE'07)* (IET, 2007), pp. 1–5.
95. M.-W. Mak and H.-B. Yu, A study of voice activity detection techniques for NIST speaker recognition evaluations, *Computer Speech & Language* **28**(1) (2014) 295–313.
96. W. M. Campbell *et al.*, Support vector machines for speaker and language recognition, *Comput. Speech Lang.* **20**(2) (2006) 210–229.
97. H. Fenglei and W. Bingxi, Text-independent speaker recognition using support vector machine, *Proc. ICII 2001-Beijing. 2001 Int. Conf. Info-tech and Info-net, 2001* (IEEE, 2001).
98. J. Cao, N. Kubota and H. Liu, A two-stage pattern matching method for speaker recognition of partner robots, in *IEEE Int. Conf. Fuzzy Systems (FUZZ), 2010* (IEEE, 2010), pp. 1–6.
99. B. Peskin, J. Navratil, J. Abramson, D. Jones, D. Klusacek, D. Reynolds and B. Xiang, Using prosodic and conversational features for high-performance speaker recognition: Report from JHU WS'02, in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, 2003 (ICASSP'03)*, Vol. 4 (IEEE, 2003), pp. IV–792.

100. T. Kinnunen *et al.*, What else is new than the hamming window? Robust MFCCs for speaker recognition via multitapering, *INTERSPEECH* (2010).
101. M. J. Alam *et al.*, Comparative evaluation of feature normalization techniques for speaker verification, in *Advances in Nonlinear Speech Processing* (Springer, NY, 2011), pp. 246–253.
102. Y. Yu, Distant talker labelling for a conferencing situation using data from a large aperture microphone array in a noisy reverberant environment, Master thesis, Brown University (2007).
103. A. Punchihewa and Z. Mohd Arshad, Voice command interpretation for robot control, in *5th Int. Conf. Automation, Robotics and Applications (ICARA), 2011* (IEEE, 2011), pp. 90–95.
104. C. Zieger and M. Omologo, Combination of clean and contaminated GMM/SVM for far-field text-independent speaker verification, *INTERSPEECH* (2008).
105. T. Yamada, L. Wang and A. Kai, Improvement of distant-talking speaker identification using bottleneck features of DNN, *INTERSPEECH* (2013).
106. N. Mubeen, A. Shahina and G. Vinoth, Combining spectral features of standard and Throat Microphones for speaker identification, in *Int. Conf. Recent Trends in Information Technology (ICRTIT), 2012* (IEEE, 2012), pp. 119–122.
107. N. A. G. H. Shindala, Investigation of distance effect on Gaussian Mixture Models in Speaker Identification, *Al-Rafadain Engineering Journal* **19**(5) (2011) 53–65.
108. R. C. Price, J. P. Willmore, W. J. J. Roberts and K. J. Zyga, Genetically optimised feedforward neural networks for speaker identification, in *Proc. Fourth Int. Conf. Knowledge-Based Intelligent Engineering Systems and Allied Technologies, 2000*, Vol. 2 (IEEE, 2000), pp. 479–482.
109. P. Kenny, P. Ouellet, N. Dehak, V. Gupta and P. Dumouchel, A study of interspeaker variability in speaker verification, *IEEE Transactions on Audio, Speech, and Language Processing* **16**(5) (2008) 980–988.
110. Y. Lei and J. H. L. Hansen, Mismatch modeling and compensation for robust speaker verification, *Speech Commun.* **53**(2) (2011) 257–268.
111. H. A. Murthy *et al.*, Robust text-independent speaker identification over telephone channels, *IEEE Trans. Speech Audio Process.* **7**(5) (1999) 554–568.
112. L. P. Wong and M. Russell, Text-dependent speaker verification under noisy conditions using parallel model combination, in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, 2001 (ICASSP'01)*, Vol. 1 (IEEE, 2001), pp. 457–460.
113. N. Dehak *et al.*, Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification, *INTERSPEECH* (2009).
114. C. Longworth, Kernel methods for text-independent speaker verification, Ph.D. thesis, Cambridge University and Christ College (2010).
115. A. Stolcke, S. Kajarekar and L. Ferrer, Nonparametric feature normalization for SVM-based speaker verification, in *IEEE Int. Conf. Acoustics, Speech and Signal Processing, 2008 (ICASSP 2008)* (IEEE, 2008), pp. 1577–1580.
116. R. Auckenthaler, M. Carey and H. Lloyd-Thomas, Score normalization for text-independent speaker verification systems, *Digital Signal Process.* **10**(1) (2000) 42–54.
117. M. Ben, R. Blouet and F. Bimbot, A Monte-Carlo method for score normalization in automatic speaker verification using Kullback-Leibler distances, in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), 2002*, Vol. 1 (IEEE, 2002), pp. I–689.
118. K. Markov and S. Nakagawa, Frame level likelihood normalization for text-independent speaker identification using Gaussian mixture models, in *Proc. Fourth Int. Conf. Spoken Language, 1996 (ICSLP 96)*, Vol. 3 (IEEE, 1996), pp. 1764–1767.

119. H. Gish and M. Schmidt, Text-independent speaker identification, *IEEE Signal Process. Maga.* **11**(4) (1994) 18–32.
120. R. Zheng, S. Zhang and B. Xu, Text-independent speaker identification using GMM-UBM and frame level likelihood normalization, in *Int. Symp. Chinese Spoken Language Processing, 2004* (IEEE, 2004), pp. 289–292.
121. H. L. Van Trees, *Optimum Array Processing* (*Detection, Estimation and Modulation Theory, Part IV*), Vol. 65 (John Wiley and Sons, New York, 2002), pp. 3185–3201.
122. S. Argentieri *et al.*, Binaural systems in robotics, in *The Technology of Binaural Listening* (Springer, NY, 2013), pp. 225–253.
123. Y. Obuchi, Mixture weight optimization for dual-microphone MFCC combination, in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005* (IEEE, 2005), pp. 325–330.
124. D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (Wiley-IEEE Press, NJ, 2006).
125. K. J. Palomäki, G. J. Brown and D. Wang, A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation, *Speech Commun.* **43**(4) (2004) 361–378.
126. M. Cooke *et al.*, Robust automatic speech recognition with missing and unreliable acoustic data, *Speech Commun.* **34**(3) (2001) 267–285.
127. J. S. Gammal and R. Goubran, Combating reverberation in speaker verification, in *Proc. IEEE Instrumentation and Measurement Technology Conference, 2005 (IMTC 2005)*, Vol. 1 (IEEE, 2005), pp. 687–690.
128. V. Hautamäki, M. Tuononen, T. Niemi-Laitinen and P. Fränti, Improving speaker verification by periodicity based voice activity detection, in *Proc. 12th Int. Conf. Speech and Computer (SPECOM'2007)* (2007), pp. 645–650.
129. J. Pohjalainen, R. Saeidi, T. Kinnunen and P. Alku, Extended weighted linear prediction (XLP) analysis of speech and its application to speaker verification in adverse conditions, in *INTERSPEECH* (2010), pp. 1477–1480.
130 T. Kinnunen, Md. J. Alam, P. Matejka, P. Kenny, J. Cernocký and D. D. O'Shaughnessy, Frequency warping and robust speaker verification: A comparison of alternative mel-scale representations, in *INTERSPEECH* (2013), pp. 3122–3126.
131. B. Manevarte, W. Ahmad and R. M. Hegde, Distant speaker verification using a combined family of MVDR estimates, in *Advances in Multimedia Information Processing–PCM 2012* (Springer, NY, 2012), pp. 628–638.
132. J. Luo, C.-C. Leung, M. Ferras and C. Barras, Parallelized factor analysis and feature normalization for automatic speaker verification, in *INTERSPEECH* (2008), pp. 1409–1412.
133. O. Buyuk and M. L. Arslan, Model selection and score normalization for text-dependent single utterance speaker verification, *Turk. J. Electr. Eng. Comput. Sci.* **20** (2012) 1277–1295.
134. R. Zheng, S. Zhang and B. Xu, A comparative study of feature and score normalization for speaker verification, *Advances in Biometrics* (Springer, NY, 2005), pp. 531–538.
135. H. Tang, Z. Chen and T. S. Huang, Comparison of algorithms for speaker identification under adverse far-field recording conditions with extremely short utterances, in *IEEE Int. Conf. Networking, Sensing and Control, 2008 (ICNSC 2008)* (IEEE, 2008), pp. 796–801.
136. S. Cumani and P. Laface, Memory and computation trade-offs for efficient i-vector extraction, *IEEE Trans. Audio, Speech Lang. Process.* **21**(5–6) (2013) 934–944.
137. D. Wang, On ideal binary mask as the computational goal of auditory scene analysis, in *Speech Separation by Humans and Machines* (Springer, NY, 2005), pp. 181–197.

138. T. May, S. van de Par and A. Kohlrausch, A probabilistic model for robust localization based on a binaural auditory front-end, *IEEE Trans. Audio, Speech Lang. Process.* **19**(1) (2011) 1–13.
139. A. A. Nugraha, K. Yamamoto and S. Nakagawa, Single-channel dereverberation by feature mapping using cascade neural networks for robust distant speaker identification and speech recognition, *EURASIP J. Audio, Speech Music Process.* **2014**(1) (2014) 13.
140. N. R. Shabtai, B. Rafaely and Y. Zigel, The effect of reverberation on the performance of cepstral mean subtraction in speaker verification, *Appl. Acoust.* **72**(2) (2011) 124–126.
141. N. R. Shabtai, B. Rafaely and Y. Zigel, Room volume classification from reverberant speech, *Workshop on Acoustics Signal Enhancement (Tel Aviv, Israel, 2010)* (IEEE, 2010).
142. N. R. Shabtai, Y. Zigel and B. Rafaely, Room volume classification from room impulse response using statistical pattern recognition and feature selection, *J. Acoust. Soc. Am.* **128** (2010) 1155.
143. N. R. Shabtai, Y. Zigel and B. Rafaely, Estimating the room volume from room impulse response via hypothesis verification approach, in *IEEE/SP 15th Workshop on Statistical Signal Processing, 2009 (SSP'09)* (IEEE, 2009), pp. 717–720.
144. N. R. Shabtai, Y. Zigel and B. Rafaely, Feature selection for room volume identification from room impulse response, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2009 (WASPAA'09)* (IEEE, 2009), pp. 249–252.
145. S. Haykin, *Adaptive Filter Theory (ise)* (2003).
146. R. Gomez, K. Nakamura, K. Nakadai, U.-H. Kim, H. G. Okuno and T. Kawahara, Hands-free human-robot communication robust to speaker's radial position, in *IEEE Int. Conf. Robotics and Automation (ICRA), 2013* (IEEE, 2013), pp. 4329–4334.
147. N. R. Shabtai, Y. Zigel and B. Rafaely, The effect of room parameters on speaker verification using reverberant speech, in *IEEE 25th Conv. Electrical and Electronics Engineers in Israel, 2008 (IEEEI 2008)* (IEEE, 2008), pp. 231–235.
148. S. Gannot and M. Moonen, Subspace methods for multimicrophone speech dereverberation, *EURASIP J. Adv. Signal Process.* **2003**(11) (1900) 1074–1090.
149. A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin *et al.*, The ICSI meeting corpus, in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, 2003 (ICASSP'03)*, Vol. 1 (IEEE, 2003), pp. I–364.
150. V. R. Algazi, R. O. Duda, D. M. Thompson and C. Avendano, The CIPIC HRTF database, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2001* (IEEE, 2001), pp. 99–102.
151. H. Steusslof, A. Waibel and R. Stiefelhagen, Computers in the human interaction loop, available at http://chil.server.de.
152. M. Lincoln, I. McCowan, J. Vepa and H. K. Maganti, The multi-channel Wall Street Journal audio visual corpus (mc-wsj-av): Specification and initial experiments, in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005* (IEEE, 2005), pp. 357–362.

Mohammad Ali Nematollahi was born in 1986 in Shiraz, Iran. He received his Bachelor of Science degree in Computer Engineering (software) in Yazd, Iran in 2008. He completed his Master's degree in Computer Engineering (software) in Islamic Azad University (IAU), Tehran, Iran in 2011. He worked as a technical expert and programmer in Ncomputing Company, Dubai, UAE from 2008 to 2010. He was also a lecturer in IAU in 2010 and 2011. He is a Ph.D. graduate in Computer and Embedded System Engineering from Universiti Putra Malaysia (UPM) in 2015. His research interests include digital signal processing, speaker recognition and digital watermarking.

Syed Abdul Rahman is a Ph.D. graduate in Electrical, Electronic and System Engineering from Universiti Kebangsaan Malaysia (UKM). He has served as a lecturer in Department of Computer Engineering and Communications System, Universiti Putra Malaysia since 2000 before promoted as a Senior Lecturer in 2006 and Associate Professor in year 2012. He also had written numerous journals which were mainly published in IEEE, WSEAS Transactions on Communication Journal, American Journal of Applied Sciences, and ASM Science Journal. His research interests include Speech Recognition, distributed system, and database.