IET Research Journals

Review Article

The Institution of Engineering and Technology

Speaker Verification with Short Utterances: A Review of Challenges, Trends and Opportunities

ISSN 1751-8644
doi: 0000000000
www.ietdl.org

Arnab Poddar[1] *Md Sahidullah*[2] *Goutam Saha*[1]

[1] *Dept of Electronics and Electrical Communication Engineering, Indian Institute of Technology Kharagpur, India*
[2] *School of Computing, University of Eastern Finland, Joensuu, Finland*
✉ *E-mail: arnabpoddar@iitkgp.ac.in, sahid@cs.uef.fi, gsaha@ece.iitkgp.ernet.in*

Abstract: Automatic speaker verification (ASV) technology now reports a reasonable level of accuracy in its applications in voice based biometric systems. However, it requires adequate amount of speech data for enrollment and verification, otherwise the performance becomes considerably degraded. For this reason, the trade-off between the convenience and security is difficult to maintain in practical scenarios. The utterance duration remains a critical issue while deploying a voice biometric system in real-world applications. A large amount of research work has been carried out to address the limited data issue within the scope of speaker verification. The advancements and research activities in mitigating the challenges due to short utterance have seen a significant rise in recent times. In this paper, we present an extensive survey of speaker verification with short utterances considering the studies from recent past and include latest research offering various solutions and analysis. The review also summarizes the major findings of the studies of duration variability problem in ASV systems. Finally, we discuss a number of possible future directions promoting further research in this field.

1 Introduction

Speech signal conveys information associated with the physiological properties of a speaker [1, 2], as it reflects the distinctive size and shape of vocal-tract, mouth, nasal cavity etc. It also contains information related to behavioral aspects of a speaker like accent, involuntary transforms of acoustic parameters etc. Hence, voice samples are often used as a biometric in real-world. Speaker recognition is the process of automatically recognizing a speaker from his/her voice samples. Speaker recognition activity can be divided into two principle tasks, speaker identification (SI) and speaker verification (SV). Speaker identification is to identify a speaker from a given set of speakers from the input speech signal [3]. Automatic speaker verification (ASV) addresses the authentication issue of a claimed identity of a person from his/her voice samples. ASV systems can broadly be categorized into text-dependent (TD) [4] and text-independent (TI) [1] types, based on the lexical content of the spoken voice. The TD-ASV requires the same lexical content for both enrollment and testing. In case of TI-ASV, it posses no restriction on the text/phonetic content of speech.

The present ASV systems are gaining acceptance for improving security in different sectors like finance, banking, surveliance, etc. [1, 2], where voices are being used as biometric [5, 6]. The state-of-the-art ASV systems exhibit satisfactory performance with adequately long speech data. However, the requirement of significant amount of speech for training or testing, especially with large intersession variability has limited the potential of its widespread practical implementations. An ASV system, in real-world, is constrained on the amount of speech data. Though the requirement in training utterance duration can somehow be taken care of by collection of adequate speech data, it is not always feasible to procure the same in verification. In case of forensic application of speaker verification systems, it is less likely to get sufficient data even for enrollment [7]. In access control type cases, average utterance length is restricted to a few seconds only [8]. Hence, it is important to take up research efforts to get reliable ASV performance in short duration condition.

Over the years, the effort to develop an ASV system suitable for real-world implementation has experienced a significant progress. Rigorous reviews on ASV systems, presented in [1, 2, 9], considered overall issues and techniques in ASV and also mentioned limited duration as one of the problems in ASV. However, in recent times, a considerable amount of research in ASV focuses on the short utterance issue, which is more challenging in practical scenarios. Hence, an extensive and systematic review can help to summarize the published research works in this area and to explore further opportunities. We discuss the issue with the short utterances in different components of ASV. This paper surveys the contemporary approaches to address the research problems related to ASV especially with short utterances. A systematic study is presented in this paper that focuses on the performance studies and analysis, trends, major challenges for the advancements of ASV in degraded and challenging conditions. The research articles published so far are categorized into different groups considering the type of work, and the methods proposed for improvements. As mentioned before, there exists a number of very useful review articles on speaker verification in general [1, 2, 9, 10]. Unlike earlier studies, this work provides extensive review of the speaker verification focusing on the most relevant short utterance problem.

The rest of the paper is organized as follows: Section 2 illustrates a brief overview of the commonly used ASV systems: GMM-UBM [11] and modern i-vector [12]. Section 3 discusses the short utterance issues and challenges in the existing ASV systems. The Section 4 presents the rigorous literature review of the published works on this topic. We summarize the findings in Section 5. We further discuss the future scopes in Section 6 and conclusion is drawn in Section 7.

2 Brief Overview of Automatic Speaker Verification

An ASV system includes three fundamental modules [1, 2]: a feature extraction unit, which transforms the speech signal in a compact form, a statistical modeling unit to characterize the extracted features, and finally a classification module to classify a test speech.

2.1 Feature extraction approaches

The state-of-the-art ASV systems uses three major types of feature extraction techniques: sub-segmental, segmental and supra-segmental analysis. Speech signals, analyzed using the frame size and shift in the range of 3-5 ms is known as *sub-segmental* analysis [13]. Studies made in [14–16] revealed that speaker-specific excitation source information captured using the sub-segmental analysis contains considerable speaker specific information.

In case of *segmental* analysis, speech is windowed with frame size and shift in the range of 10-30 ms to extract the speaker information mainly characterizing the vocal tract. The speaker-specific vocal tract information can be assumed to be stationary for practical analyses and processing when frames of size and shift is kept in the range of 10-30 ms. Studies conducted in [17–23] used segmental features to extract the vocal tract information for verification of speakers.

In *supra-segmental* feature extraction, speech is truncated using the frame size and shift in the range of 100-300 ms. Primarily, this technique is utilized to analyze and extract characteristics of behavioral traits of the speaker. These incorporates the information of word duration, intonation, speaking rate, accent, etc. The relevant research work done in [14, 24–26] demonstrated that some behavioral traits can be captured using supra-segmental analysis which proved to be effective for speaker verification.

State-of-the-art ASV systems mostly use short-term spectral features. Mel-frequency cepstral coefficients (MFCC), perceptual linear prediction (PLP) and linear predictive cepstral coefficients (LPCC) are widely used feature extraction techniques due to their considerable performance and lower-computational complexity [1, 9, 19, 23, 27–29].

2.2 Classification approaches

The major advancements in speaker verification research is due to the improvements in classifier domain. The primitive speaker verification systems were developed using vector quantization (VQ), dynamic time warping (DTW) [30] approach. Later on, with the introduction of Gaussian mixture model (GMM) [31], ASV research has evolved in past two decades with more focus on channel compensation, data variability, etc. GMM with the universal background model (UBM) [11] was proposed with significant improvement over independently trained GMM using maximum likelihood approach [31]. The latent variable approach has introduced another new paradigm in ASV technology. For example, factor analysis (FA) based approaches were proposed to model the inter-session variability in the context of GMM *supervector* [32]. Motivated by the success of Joint FA (JFA), i.e, speaker factors directly as features for classification, Dehak *et al.* introduced single total-variability subspace based modeling of the speakers, unlike separate subspaces for speakers and channels in JFA [12]. Recent speaker verification technology focused on *total variability* modeling, also known as *i-vector*. The i-vector space is further modeled using a separate speaker and channel dependent subspaces with Gaussian probabilistic LDA (GPLDA) [33], and this approach efficiently handles the intersession variability. Current ASV technology uses this i-vector approach which provides an elegant framework to obtain a fixed dimensional representation of variable length speech utterances. In recent times, deep learning based approaches has attracted much attention and caused extensive interests in various domains[34]. For speaker verification, the studies applied DNN models trained for speech recognition to build UBM like acoustic models, so that rich information in phones can be employed to develop more efficient background models [35–37]. DNNs are also successfully implemented to extract features of speaker information [38, 39].

Here, we briefly review two widely used modeling techniques for speaker recognition: classical GMM-UBM and state-of-the-art i-vector, for the completeness of the study. A schematic block diagram of i-vector and GMM-UBM based ASV system is presented in Fig 1.

2.3 GMM-UBM system

The GMM-UBM system is a straightforward generative approach for ASV task. In this framework, training phase is preceded by estimation of a speaker-independent universal background model (UBM), using a sufficiently large speech data of several hours from multiple sources [11]. The UBM is represented as

$\boldsymbol{\lambda}_{\text{UBM}} = \{w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{i=1}^{C}$,

where C is the number of Gaussian components, w_i is the prior or weight of the i-th Gaussian component, $\boldsymbol{\mu}_i$ represents the mean and $\boldsymbol{\Sigma}_i$ is the co-variance matrix of the i-th Gaussian component. The parameter w_i satisfies the constrain $\sum_{i=1}^{C} w_i = 1$. Each speaker is represented as a GMM derived by *maximum-a-posteriori* (MAP) adaptation from UBM. For this purpose first, sufficient statistics of the features from speaker's enrollment utterances are computed. Then relevance MAP approach is used to estimate the weights, means and covariances of the target speaker model, λ_{target}. During test or verification, average log-likelihood ratio $\Lambda(\mathbf{X})$ is estimated using feature vectors from T speech frames of test utterance $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ against both target speaker model and the UBM. We can express this as,

$$\Lambda(\mathbf{X}) = \log\ p(\mathbf{X}|\boldsymbol{\lambda}_{\text{target}}) - \log\ p(\mathbf{X}|\boldsymbol{\lambda}_{\text{UBM}}) \tag{1}$$

Where $p(\mathbf{X}|\boldsymbol{\lambda})$ is the likelihood of test feature vector computed for the model $\boldsymbol{\lambda}$. Finally, a decision threshold is used to determine whether a test utterance will be accepted or rejected.

2.4 i-vector system

The ASV systems based on i-vector represent the high dimensional GMM supervector in a total variability (TV) space which reduces the supervector into low dimensional factors [12]. In TV space, GMM supervector i.e., the concatenated means of GMM components, is projected as

$$\mathbf{M} = \mathbf{m} + \boldsymbol{\Phi}\mathbf{y} \tag{2}$$

where $\boldsymbol{\Phi}$ is a low-rank factor loading matrix and $\mathbf{m}$ is channel and speaker independent supervector, (same as GMM-UBM supervector). $\mathbf{M}$ represents the supervector of the speech utterance with feature vectors $\mathbf{X}$. Total variability factors $\mathbf{y}$ are assumed to be normally distributed $[\mathcal{N}(0, \mathbf{I})]$ random variables.

Channel compensation in i-vector subspace is accomplished through probabilistic LDA (PLDA) modeling approach. For this purpose deferent variants of PLDA exists [41] and simplified Gaussian PLDA (GPLDA) approach is commonly used [33]. GPLDA approach models the inter-speaker variability of total variability subspace using a full co-variance with residual. The generative model for s-th speaker and j-th recording in projected variability space is given by

$$\boldsymbol{y}_{s,j} = \boldsymbol{\eta} + \boldsymbol{\Psi}\boldsymbol{z}_{s,j} + \boldsymbol{\epsilon}_{s,j} \tag{3}$$

where $\boldsymbol{\eta}$ is the speaker and channel independent i-vector, $\boldsymbol{\Psi}$ represents eigenvoice subspace, $\boldsymbol{\epsilon}$ represents the variability not captured by the latent variables and $\mathbf{z}_{s,j}$ is the vector of latent factors. GPLDA based i-vector system uses log likelihood ratio as verification score [33]. For a projected enrollment and verification i-vector $\mathbf{z}_{\text{target}}$ and $\mathbf{z}_{\text{test}}$ respectively, the log-likelihood ratio $\Lambda_{\text{GPLDA}}(\mathbf{X})$ can be calculated as follows,

$$\Lambda_{\text{GPLDA}}(\mathbf{X}) = \log\ \frac{p(\mathbf{z}_{\text{target}}, \mathbf{z}_{\text{test}}|H_1)}{p(\mathbf{z}_{\text{target}}|H_0)\ p(\mathbf{z}_{\text{test}}|H_0)} \tag{4}$$

where H_1 is the hypothesis for the i-vectors belonging to the same speaker and H_0 is the hypothesis for the i-vectors belonging to different speakers.

3 Problems with Short Segments

The state-of-the-art ASV systems show considerable verification accuracy but with a large amount of speech data. Typically, the NIST

This article has been accepted for publication in a future issue of this journal, but has not been fully edited.
Content may change prior to final publication in an issue of the journal. To cite the paper please use the doi provided on the Digital Library page.

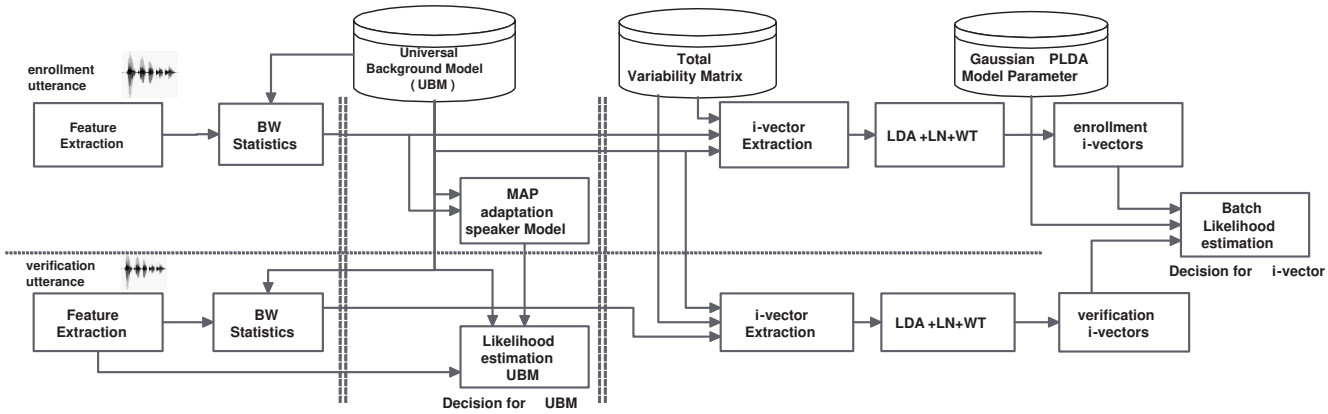

Fig. 1: Block representation of state-of-the-art i-vector and GMM-UBM based ASV system [40]

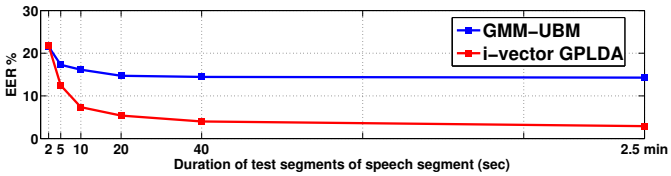

Fig. 2: Performance of speaker verification accuracy in terms of EER (%) for test with segments of different length using core-condition of NIST SRE 2010 corpus. The short test segments are created by truncating the original speech data. The training duration is fixed at $\sim$ 2.5 min.

speaker recognition evaluation corpus (SRE), which are widely used in ASV research, consists speech segments of nearly $\sim$2.5 minute duration. A few NIST SRE corpora like NIST SRE 2010, NIST SRE 2008 introduced specific tasks where short duration of nearly 10 sec segments were used [43, 44]. There is no standard definition of short duration in ASV. However, We observed that most of the published literatures considered segments of duration 5-10 sec as short utterances for experimental evaluation and analysis [7, 45].

The performance of the ASV systems degrades drastically with the reduction in speech duration in training or testing. Figure 2 shows the performance in terms of *equal error rate* (EER) for GMM-UBM and i-vector-GPLDA system to show the effect of duration of test segments. The performance of i-vector based system drops from 3.48% to 22.09% EER when the verification segment duration was shortened from $\sim$2.5 min to 2 seconds [40]. The main challenge in achieving high performance with short duration is the increase in intra-speaker variability of estimated parameters. This is due to the dependency of the parameters on the lexical content and for unconstrained setup (i.e., for text-independent), they are considerably different over different short segments [8].

The faulty recordings also reduce a major part of voice for glitches and chirps, thereby leaving a little amount of intelligible speech for further processing. In i-vector based ASV systems, utterance length is directly related to the uncertainty of i-vector point estimation [46, 47]. With a reduction in speech duration, uncertainty of i-vector estimation increases and vice-versa. The uncertainty can be estimated by computing the inverse of the trace of covariance matrix of the i-vector posterior distribution. When speech duration is longer, the covariance of the posterior distribution becomes smaller [48]. On the other hand, when utterance is short, the covariance becomes larger. As a consequence, the i-vectors are poorly estimated for short segments.

In order to observe the impact of duration on state-of-the-art i-vector system, we have performed analysis on i-vectors for different duration conditions, e.g., 2 sec, 10 sec and 20 sec in Fig. 3. Scatter plots of principal component projected i-vectors from truncated speech segments of 2 speakers from NIST SRE 2008 corpora are presented in Fig. 3. Short utterances show higher variability, and this variability decreases with the increase in utterance duration. Moreover, the inter-speaker variability is reduced in short duration. In a recent study, we have shown that the variability in i-vectors is due to the variability in estimated sufficient statistics required for i-vector computation [49].

The source of variability seems to be largely related to the linguistic content of the short section of the utterance used for i-vector extraction [22]. In traditional longer utterance i-vector extraction, this linguistic variation is averaged over a large amount of speech content, and the local variations is ignored. But in short duration condition, it plays an important role by increasing the intra-speaker variability [22]. To study the consequences of short duration utterances on the phonetic distribution, the work in [42] collected telephone recordings from the SRE'04,05 and 06 corpora and English phonemes were detected for different duration *viz.* 2, 5, 10, 20 and 40 seconds, using the work [50]. Fig. 5 shows that average number of unique phonemes reduces exponentially with duration of the segment. This partially indicated one of the causes of performance degradation in an exponential manner [51]. Histograms of phone from a single utterance in its full duration, and their truncated versions are depicted in Fig. 4. The effect of number of samples of unique phonemes and presence of unique phonemes in train and test phase of speaker verification requires a deeper investigation [42].

4 Research in ASV on Short Utterances

The short utterance has been an open challenge to the ASV research community for decades. There has been numerous attempts to mitigate the issue. The relevant works concentrated on different aspects of ASV e.g, feature extraction techniques, intermediate parameter estimation, speaker model generation, score normalization techniques etc. to compensate the duration variability issue. The research dealing with the feature extraction, classification sub-systems of ASV to mitigate the issue of short utterance are presented hereafter. A diagrammatic representation of short utterance research is presented in the Fig. 6.

4.1 Feature extraction from short utterance

The feature extraction techniques are supposed to characterize the physiological aspects of speech production system including vocal tract and source. The ASV systems dealing with short duration challenge mostly concentrate on sub-segmental cepstral features. Some of the work also analyze the wavelet-based approaches. Here, we discuss the research efforts that deal with short duration issue in ASV from feature extraction perspective.

The work in [52] proposed an algorithm for noise separation based on constrained non-negative matrix factorization (CNMF), especially in limited duration condition. The speech data were segregated to high and low quality classes using differences detection

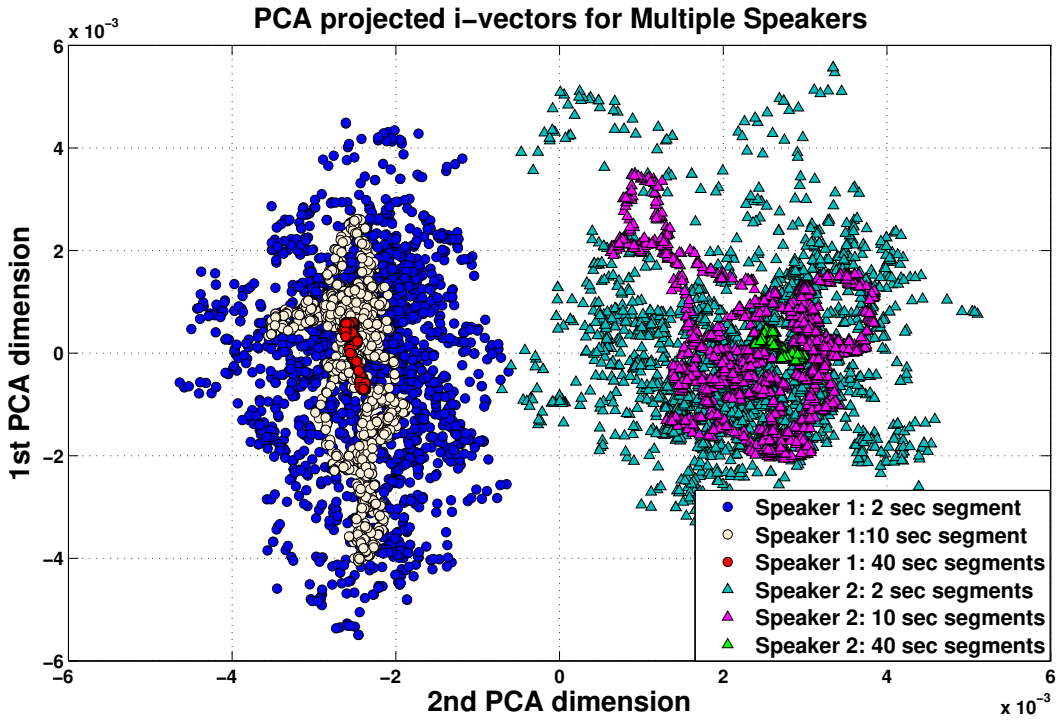

Fig. 3: 2-D scatter plot of principal component analysis (PCA) projected i-vectors for different truncated segments of 2 sec, 10 sec and 20 sec for two speakers from NIST SRE 2008. The truncated segments are created from long duration segments of approximately ∼2.5 min.

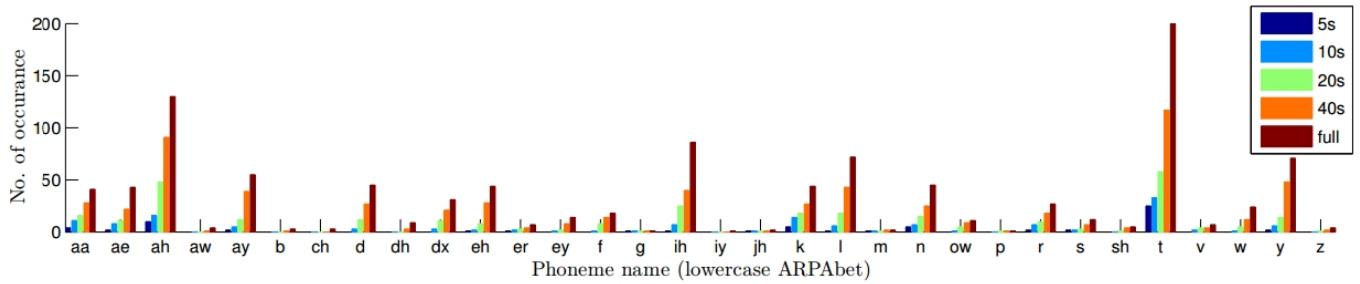

Fig. 4: Histogram of phonemes detected from an utterances in five different truncated conditions: 5 s, 10 s, 20 s, 40 s and full duration [42].

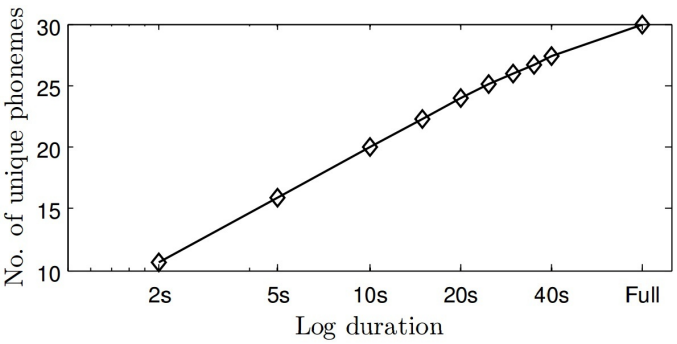

Fig. 5: Number of unique phonemes detected from varying duration utterances. The figure is obtained by averaging over 19167 utterances from NIST SRE'04,05,06 gender-mixed telephone data [42].

and discrimination algorithm (DDADA), amalgamating groups of features with GMM-UBM two-stage classification model to utilize limited information maximally.

Motivated by the analysis of multiple different scales, notably in short duration, the work in [53] proposed a multi-resolution time frequency feature (MRTF) extraction technique. The technique in [53] performed a multi-scaled 2-D DCT operation on the time frequency spectrogram matrix and then selecting and combining to the final multi-scaled transformed elements. In contrast to the traditional MFCC features, the proposed feature extraction technique can better use multi-resolution time-frequency information which could be useful for short utterance condition.

The paper [54] fused features based on amplitude spectrum, phase spectrum and combined amplitude phase spectrum especially in short utterance. The MFCC, multi-taper MFCC (MMFCC) and Linear frequency cepstral coefficients (LFCC) were used in the amplitude-based feature category. The phase-based features, used here, are: LP residual phase cepstra (LPRPC), stabilized weighted LP-GDCC (SWLP-GDCC), linear prediction-group delay cepstral coefficients (LP-GDCC), modified group delay cepstral coefficients (MGDCC). Combined amplitude-phase-based feature category includes Product spectrum-based MFCC (PS-MFCC). The combination of these features showed encouraging performance in short utterance. The work in [55–57] explored mel power difference of spectrum in sub-band (M-PDSS), residual MFCC, and discrete cosine transform (DCT) of the integrated linear prediction (LP) residual (DCTILP) for ASV under constrained duration. The

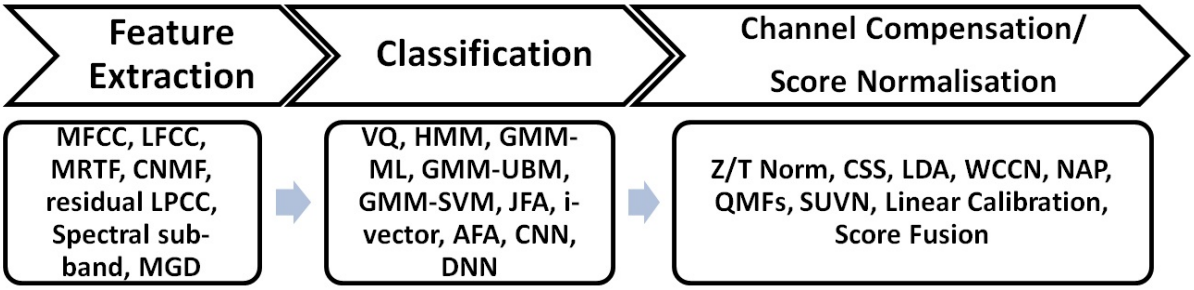

Fig. 6: Diagrammatic representation of different methods used in three sub-system levels of ASV to mitigate the problem of short utterance.

aforementioned three source features were found to capture various attributes of source information, such as, periodicity, smoothed spectrum information, and shape of the glottal signal, respectively.

4.2 VQ on short utterance

Simple vector quantization (VQ) models, also termed as centroid model [58, 59], are often used as computational speed-up techniques [60] and lightweight practical implementations. Here, we briefly discuss the modification on VQ model to address the short duration challenge.

In the work [61], a statistical model based on VQ was developed to represent the multi-modal distribution of the features. However, the study in [62] introduced a combined speaker recognition system based on VQ discrete hidden Markov model (HMM), which was designed to cope with limited number of training repetitions of vocabulary of spoken computer commands. The report [63] compared ASV systems based on a VQ distortion method and ergodic HMM, emphasizing robustness against duration variability. It subsequently showed that a continuous ergodic HMM is considerably robust as a VQ distortion method in longer speech duration. However, a continuous ergodic HMM outperformed a discrete ergodic HMM [63] in limited data. In VQ based ASV systems, the minimum of overall average distortion rule could be a criterion to classify a given speech segment to a speaker model, i.e, the codebook [64]. In the work [64] a decision rule had been introduced based on fuzzy c-means clustering. A bunch of membership functions related with vectors of codebooks were used to characterize as discriminant functions. Moreover, the maximum overall average membership function rule was used more effectively in short duration [64]. A score normalization and selection strategy was introduced to handle the poor performance with short utterances in VQ based ASV system in [65]. This work described a normalization of classification scores that depends on the means and variances of scores of short samples, matched to different models of many speakers. The selection procedure discarded a portion of a speech sample of poor speaker discrimination ability[65].

4.3 GMM on short utterance

Merlin *et al.* proposed an explicit feature space for speaker information representation to overcome the omnipresent intra-speaker variability coming from imbalance in phonetic structure of utterances [66]. The transformed space is believed to lead to a more stable model estimates due to less ambiguity and variability with limited enrollment data [66]. The work in [67] presented approach for mathematical modeling of duration patterns. In this approach, each word and/or phone was represented by a feature vector consisting of either the duration of individual phonemes making up the word, or the states of HMM making up the phonemes.

However, Larcher *et al.* criticized the GMM-UBM system targeting its incompetency in mobile applications for its larger data requirements [68]. Their approach incorporated the information of temporal structures (i.e. word dependency) and also multi-modal system (video) to compensate limited duration. Vogt *et al.* also used a technique based on factor analysis to develop subspace models that supposed to work efficiently with short duration and enabled to be combined seamlessly with the optimal speaker model using GMM-UBM with sufficient data [69]. The subsequent work suggested to estimate intervals of confidence for ASV performance scores, leading towards considerable accuracy with only 2-10 seconds of evaluation data [70, 71]. The work in [72] presented the dimension-decoupled GMM approach to handle short (enrollment and test) utterances in ASV. For limited data, the DD-GMM came up with more reliable results. Moreover it is computationally efficient irrespective of speech duration. This approach is seamlessly integrable with other approaches, allowing synergetic effects, and can be implemented directly in any form of GMM. The paper [73] attempted to study the influence of speech duration is evaluated in text-independent GMM and SVM framework. It is shown that the optimal frame selection exhibits a dependency on overall duration. The paper [74] highlighted the relevance of a well-tuned speech detection front-end. It considered a well-established GMM as well as SVM (on GMM mean super-vectors), showing their limitations and alternatives simultaneously. Specifically, the benefits of eigen-voice modeling in context of short duration condition was focused. Subsequently, it demonstrated fusion potential between the presented techniques and significant gains when compared to the classical standalone GMM.

The paper in [75] systematically analyzed performance measures in accordance with the duration of verification utterances used for background, score normalization and session compensation training cohorts. This work highlighted the importance of matching the speech duration of utterances in cohorts to the expected evaluation conditions. It showed the ASV performance to be sensitive particularly to the speech length in the dataset of the background. Moreover, the work in [75] found that nuisance attribute projection (NAP) approach, used for session compensation often degrades performance over other methods when speech duration remains limited.

The work in [76] subsequently introduced an artificial feature addition method for speaker identification in limited data. In the work [77], a Fisher-voice based feature fusion strategy incorporating PCA and LDA was utilized, where several commonly used features, such as MFCC, PLAR and LPCC, were concatenated, and subsequently projected onto a subspace with lower-dimensionality .

In context of mobile devices, ASV engines are susceptible to suffer from limited computing resources and ergonomic constraints. A GMM-UBM extension was prescribed in [78] to compensate the situations characterized by constrained amount of enrollment data and computation facility, typically available on hand-held mobile devices. The key contribution was influenced from the idea of incorporation of temporal structure information of speech using pass-phrases customized by the client and new Markov model structures in addition to it. Moreover, additional temporal information was utilized to increase discrimination with Viterbi decoding.

The work in [79] proposed a multi-scale kernel (MSK) learning approach to address the problem in GMM-SVM framework. It constructs a series of kernels with different scales, and combine them through multiple kernel learning (MKL) optimization. The robustness and scalability of the system can be enhanced with this approach.

Table 1 Summary of previous works on i-vector speaker verification with short utterances.

Year	#Ref.	Work Done	Methodology
2011	[45]	Performance Evaluation	JFA, i-vector CSS, WCCN, LDA, NAP, SDNAP, GPLDA
	[7]	Calibration Evaluation	Linear calibration, cosine kernel, normalized cosine kernel
2012	[80]	Performance Evaluation	Inclusion of short utterances in development data-set.
	[51]	Performance Evaluation	An ad-hoc fusion system of different total variability spaces.
	[81]	Evaluation of Phoneme effects	Adding phonetic information, WCCN and eigen factor radial (EFR)
	[82]	Adding of Phonetic information	Vowel Categories (VC's), Universal Background VC Models (UBVCM)
	[83]	Adding of Syllable information	Syllable Categories, Universal Background Syllable Models
2013	[42]	Analysis on phoneme distribution	Score calibration with log duration as QMF, synthetic i-vectors.
	[84]	Analysis of phonetic content	TD-ASV, Multiple enrollment, Used speaker and phonetic content.
	[85]	Analysis on Confusion errors	Finding speaker-specific phonemes, formulate text using unique phonemes.
	[21]	Analysis on Score Calibration	QMFs, Stacked Scores, Shared Scaling, Extrapolation
	[86]	Performance Analysis	Total Variability, PLDA
	[87]	Source and Utterance-Dur. Norm.	SUN-LDA, LDA, WCCN, CSS, SN-LDA
	[46]	BW statistics estimation	Minimax [88] strategy to estimate the BW statistics (zeroth and 1^{st} order)
2014	[22]	Source and Utterance-Var. Norm.	LDA, SUN-LDA, SNLDA, SUVN, WCCN, GPLDA
	[89]	Feature-level phone normalization	phone/speaker adaptive training (PAT/SAT), constrained MLLR
	[90]	Combining Source and System info.	DCT of the integrated LP residual (DCTILPR), Score Fusion
2015	[54]	Amplitude and Phase-based features	LFCC, M-MFCC, MGDF, All pole GDF, LP phase-residual
	[91]	Evaluation of Feature Dimensionality	Feature Dimension reduction, Discrete Karhunen-Love transform (DKLT)
	[40]	Performance Comparison	GMM-UBM, i-vector GPLDA
	[56]	Parallel system based on source feature	M-PDSS, DCTILPR, Score Fusion
	[92]	i-vector subspace projection	Modified-prior PLDA, Score Calibration, QMF
	[93]	Calibration and quality of speech signal	QMFs from duration + SNR, Stacked, Matched/Mismatched calibration
2016	[94]	Phonetic match between train and test	WCCN, Eigen Factor Radial (EFR), interactive voice response system (IVRS)
	[95]	Factor Analysis on i-vector domain.	Acoustic Factor Analysis (AFA), WCCN, LDA, GPLDA, Score level Fusion
	[96]	Phonetic content compensation	Max. Likelihood-acoustic factor analysis (ML-AFA), SUVN, Score Fusion
	[97]	Phonetic Analysis	Modeling speech unit classes
	[98]	Normalize BW statistics	Compensation for feature sparsity in BW statistics
	[99]	Bootstrapped i-vectors	Truncate from test segment and integrating speaker similarities
2017	[100]	Development data with short utterance	WCCN-LDA, SN-LDA, SN-WLDA, GPLDA
	[101]	inter/intra-speaker variability	A transform to map i-vectors onto a duration invariant latent subspace
	[102]	i-vector length normalization	DNN based length normalization of i-vectors using principal components

4.4 i-vector on short utterance

The i-vectors, whether through the simplistic cosine similarity scoring (CSS), or more advanced length-normalization technique GPLDA are regarded as the de-facto standard for modern speaker verification research. A summary of major researches on short utterance with i-vector based ASV system is presented in Table 1. Some results of comparison and advancements on i-vector based ASV techniques in short-duration condition are presented in Table 2 for better understanding.

4.4.1 ***Analytical studies****:* Here we attempt to present the researches which studied and investigated the short duration problem of i-vector based ASV system from different perspectives.

The work in [45] investigated ASV performance when various intersession variability compensation techniques such as within class covariance normalization (WCCN), LDA and NAP were used in conjunction with i-vectors, including a short investigation of scatter-difference NAP (SDNAP). Moreover, it compared the above combination of methods with GPLDA [33].

The report in [7] conducted a systematic performance study on forensic applications. The classification performance and calibration cost of the i-vector system was evaluated along with the significance of normalization techniques in the cosine kernel. The report [7] highlighted the normalization of the cosine kernel which provided improved performance metrics across different speech durations compared to that of the unnormalized kernel.

The work in [51] explored i-vector based ASV systems to train development parameters when target speakers were trained and tested on mismatched speech durations. It was found that short speech segments are optimal for training system model parameters when target speakers are enrolled and tested on short segments. However, in case of target speakers training on long speech segments and tested on either full (i.e. long) or short speech segments, it is more appropriate to use long speech segments for training system model parameters.

A wide analysis on short utterances was conducted in [86] which explored the limitations of the i-vector systems especially in short duration condition. However, it showed that due to the final dimensionality reduction, dimension of the i-vector systems designed with complex or heavy i-vector extractors (high number of Gaussian, i-vector dimension) are likely to provide advantages over complementary lighter ones.

The work in [40] rigorously presented a comparative study of different ASV systems with utterance duration variability. The ASV system, considered here, were based on i-vector and classical GMM-UBM. This study revealed that the relative improvement of i-vector system over GMM-UBM drastically degraded with the reduction in test utterance duration. It was noticed that for the speaker models, trained with longer training data, simplistic GMM-UBM system performed better over i-vector system in very short verification segments.

4.4.2 ***i-vector estimation and normalization****:* Here we have presented the research works which concentrated on modifications and analysis of core i-vector estimation and normalization approach. Some of results of different methods in this context are presented in Table 2.

In ASV, when the number of the feature vectors is relatively small, uncertainty in the representation of i-vector as a point estimate of the linear-Gaussian model is considerably tricky [46]. It is evident that the zeroth and first order sufficient BW statistics, given the hyper-parameters, entirely represents the extracted speaker dependent i-vectors. The study in [46] introduced a minimax strategy to estimate the intermediate BW sufficient statistics efficiently in order to enhance the robustness of i-vector based speaker model.

The work in [22] identified the phonetic content mismatch as an additional source. The work introduced a short utterance variance normalization (SUVN) strategy and a short utterance variance (SUV) modeling approach to compensate the introduced session and duration variability at the i-vector feature level. A systematic combination of SUVN with LDA and SN-LDA was further used effectively. Moreover, an alternative approach was introduced using PLDA to directly model the SUV which showed that for the combination of mentioned techniques, the information of utterance variation needs to be supplemented additionally to full-length i-vectors for modeling of PLDA.

The work in [90] analyzed the PLDA modeling with limited development data. The work scrutinized the effectiveness of the median as the central tendency of a speaker's i-vector representation, and the effectiveness of weighted discriminative methods on the performance of length normalized GPLDA systems. The presented studies showed that the median (using a median fisher discriminator (MFD)) obtained a better representation of speakers when the number of representative i-vectors available during development was lowered. Furthermore, usage of the pairwise weighting approach in weighted LDA and weighted MFD yielded more efficient performance.

In [54], the authors proposed to combine amplitude and phase based features for improving speaker verification with short utterances for text-dependent speaker verification. The work presented in [91] demonstrated that reduction of the feature dimensions with the discrete Karhunen-Love transform (DKLT), considerably helped improving performance over the baseline MFCC features. Moreover, particularly short length speech frames, i.e, with utterance length less than 1 sec, the performance of truncated DKLT representation outperformed conventional MFCC.

Subsequently, in [100] a number of techniques, including source-normalized weighted LDA (SN-WLDA) projections and utterance partitioning are implemented. They found that when development of statistical model parameters are only restricted to short utterance data, GPLDA system achieves best performance with a relatively low UBM components. This in turn, significantly reduced the computational complexity of ASV system. The work proposed a simplistic utterance-partitioning technique successfully with the conjecture that the improvement was due to apparent enhancement in the number of sessions by partitioning technique which accompanied better estimation of GPLDA parameters.

The study in [95] illustrated that combination of two methods, acoustic factor analysis (AFA) [103] and i-vector [12]) can lead to a much upgraded ASV system to mitigate the issue of short utterances. MFCC features are projected onto a lower dimensional subspace using factor analysis based in the AFA technique. The AFA system alone outperformed the i-vector based system in limited data condition. Further improvement using score level fusion of the two systems while having more weightage towards AFA in limited verification data condition. A different AFA approach is also used in [96]. An i-vector based SV system using maximum likelihood - acoustic factor analysis (ML-AFA) technique for speaker modeling was introduced in the work. Furthermore, it used short utterance variance normalization (SUVN) technique to enhance the phonetic content of extracted i-vectors in short utterances.

The work in [98] analytically and experimentally demonstrated, that feature imbalance sparsity emphatically persists in short utterances. The work presented an improved i-vector extraction algorithm with a systematic analysis of adaptive first-order BW statistics (AFSA). The algorithm attempted to compensate the deviation from first-order BW statistics caused by feature sparsity and imbalance in short utterance. A compensation technique was introduced in [101] to normalize the distribution mismatch caused by duration variation in the i-vector space. The mentioned approach involves the use of two factor analyzers, tied together, to share latent variables for a given speaker as the underlying generative model of the i-vector space.

The work in [102] proposed a method to transform short-utterance feature vectors to adequate vectors using deep neural network architecture, which compensate the variability appeared in short utterances. For the dimensionality reduction of search space, the principal components are applied from the residual vectors between every long utterance i-vector in a development set and its truncated short segment i-vector version. An i-vector of the network is transformed by linear combination of these directions.

4.4.3 ***Phonetic Analysis****:* The work in [81] showed that the ASV performance of state-of-the-art i-vector technique can considerably be enhanced using the information of the underlying phonetics. Moreover, it showed the potential of further improvements by taking opportunity to use phonetic information in the normalization stage. The study systematically compared two score normalization methods, eigen factor radial (EFR) and WCCN, both depending on parameters estimated with same development data. The comparison suggested that WCCN is more efficient to data mismatch but less effective than EFR when the development speech utterances have more similarity with the verification speech.

Table 2 Results and comparison of i-vector based ASV techniques and advancements in short-duration condition.

#Ref.	Modeling Methodology	Feature	Database (SRE)	Task	Length Train-Test	EER [%]	min DCF
[45]	JFA	MFCC	NIST '08	short2-short3	10s - 10s	21.17	0.0738
	TV+LDA+WCCN					21.56	0.0741
	TV+SDNAP+WCCN					20.84	0.0737
	TV+GPLDA					20.34	0.0762
[7]	Normalized i-vector	MFCC	NIST '10	core-core	10s - 10s	14.68	0.063
[80]	TV+GPLDA matched dev. data	MFCC	NIST '08	Short2-short3	10s - 10s	16.04	0.0679
	TV+HTPLDA matched dev. data					13.67	0.0639
[51]	TV+GPLDA (dev data: 10s segments)	MFCC	NIST '08	core	10s - 10s	10.70	0.0518
	TV+GPLDA (dev data: full segments)					11.77	0.0657
	TV+GPLDA (dev data: 10s+full segments)					9.79	0.0462
	TV+GPLDA (Fusion: 10s, full, 10s+full)					8.19	0.0442
[46]	TV+GPLDA (max. likelihood)	MFCC	NIST '10	8conv-10sec	Long-10s	9.89	-
	TV+GPLDA (minimax)					7.99	-
[90]	TV+GPLDA	MFCC	NIST '08	short2-short3	10s - 10s	15.07	0.0673
	TV+WCCN+GPLDA					14.99	0.0674
	TV+WCCN+LDA+GPLDA					15.80	0.0664
	TV+WCCN+SNLDA+GPLDA					15.40	0.0661
	TV+SUVN[LDA]+GPLDA					14.75	0.0618
	TV+SUVN[SNLDA]+GPLDA					14.73	0.0620
[92]	TV+GPLDA	MFCC	NIST '10	core-core	Long - 10s	10.4	0.0438
	TV+GPLDA (Modified Prior)					9.9	0.0464
[40]	TV+GPLDA	MFCC	NIST '08	short2-short3	10s - 10s	13.47	0.0635
	GMM-UBM					16.62	0.0700
[94]	TV+LDA+WCCN	MFCC	NIST '03	evaluation plan	Long - 10s	5.81	0.1090
		MPDSS+MFCC				5.56	0.1048
		RMFCC+MFCC				5.78	0.1087
		DCTILPR+MFCC				5.33	0.0971
[95]	TV+LDA+WCCN	MFCC	NIST '03	evaluation plan	Long - 10s	5.81	0.1090
	AFA+LDA+WCCN					4.92	0.0882
	Fusion: AFA, TV					4.29	0.0802

The paper [82] introduced an approach of using vowel categories (VC). After recognizing and extracting the phonemes, extracted vowels are divided into VC's to generate Universal Background VC Models (UBVCM) for each VC. A similar approach to GMM-UBM approach was utilized for enrollment and verification. It was shown that vowels contained significant speaker discriminative information, which remained placid when vowels categorization was deployed. A similar approach but with syllable category based speaker verification in short utterance was conducted in [83].

The work in [42] analyzed the effect of duration variation on distributions of phonemes present in speech utterances and corresponding i-vector length. It demonstrated that, as utterance length was reduced, the i-vector length and the number of unique phonemes approached zero in a non-linear and logarithmic fashion, respectively. Assuming duration variability as an additive noise in the i-vector space, it introduced experimentally verified in publicly available corpus three different compensation strategies given hereafter [42]:

- multi-duration training in PLDA
- score calibration with log of duration as a Quality Measure Function (QMF)
- PLDA model training with artificially synthesized i-vectors in short duration.

The work in the paper [84] evaluated the performance of i-vector with PLDA in a text-dependent environment. It showed that the use of class definitions of both speaker and phonetic content can significantly improve ASV system performance. Moreover, the study rigorously compared four score computing models with multiple enrollment segments and showed that intrinsic scoring of PLDA yielded enhanced performance in short duration. This study suggested that the optimization of scoring regime can help improve the system performance especially in case of short utterance [84]. The work in [85] introduced a strategy to minimize the confusion errors, by finding speaker-specific phonetic distribution and subsequently generate a text using the subset of unique phonemes.

The work in [94] projected the significance of phonetic similarity between test and train segment for a TI environment under short utterance condition. The framework recommended to implement a speaker model with constrained text estimated using limited data of around 10 sec. The similar content of phonemes is supposedly spoken by the end-user during authentication. It showed that the text constrained model based topology worked far efficiently especially in limited duration. Moreover, it was shown that presence of similar phonetic content of the verification session in enrollment, helped outperform the baseline system.

The work in [97] argued that the short duration problem in ASV can mostly be ascribed to the mismatched prior distributions of the speech data used for training. The paper in [97] introduced an approach that distributed speech signals into a multitude of acoustic subregions, defined by models speakers and phonetic units within the subregions. To avoid data sparsity, an approach, driven by data, was used for clustering speech unit classes, based on estimated models of subregions. Furthermore, it successfully implemented a synthetic modeling approach based on MLLR inclusively with no-data speech unit classes.

4.4.4 ***Score Calibration****:* The paper [21] investigated the effect of calibration on short duration condition. A calibration approach was introduced using quality measure functions (QMFs) to incorporate duration information in final verification score. In this work extensive experiments highlighted the importance of considering the metrics of speech quality like duration for calibrating the scores of ASV. Incorporation of additional information via quality measure could significantly help improve the ASV performance in short utterance. In extension to this work, the paper [93] investigated the consequences of short duration and noisy speech. The work proposed simple speech Quality Measure Functions (QMFs) of duration and measured signal-to-noise-ratio (SNR).

5 Summary

In this section, we summarize the major achievements and contributions made so far to address the short utterance challenge in ASV.

- Major researches attempted to compensate the effect of short utterance by reducing the mismatch of duration occurred in statistical modeling of speakers [12, 39, 90, 97].
- The state-of-the-art i-vector and GMM-UBM based ASV systems use sufficient statistics of feature vectors computed from speaker-independent feature space. Improved estimation of this statistics and its alternative estimation considering parameters such as duration have been done [46].
- The existing work mostly concentrated with publicly available databases (NIST SRE, WSJ etc.), which consist of more than 2.5 minutes of speech. Though some database have evaluation condition with speaker verification task in short utterance, most of the researches work is done by truncating the longer version of speech signal of core evaluation condition [21, 22, 45, 46, 55].
- Studies have shown that in very short duration cases, the classical GMM-UBM based approach worked better with respect to the modern i-vector based approach [40]. Fusion of multiple classifiers yielded considerable improvements over the standalone approaches [104].
- Research in short utterance problem in ASV has seen efforts to accommodate phonetic distribution for speaker modeling [57, 97]. This approach resulted significant improvements in the relevant problem.
- Calibration and normalization of classification scores successfully attempted in many work to reduce the duration variability effect in short utterances [22, 93, 100]. Inclusion of additional information such as quality or intelligibility metric of speech played an important role [21, 105]. In this regard, variability modeling as a compensation strategy seems to be useful [22].

The extensive discussions in Section 4 show that the published works in speaker verification with short duration address one of shortcomings as a results of limited speech. The works presented in Section 4, use different speech corpus and different ASV subconditions for analysis and evaluation of proposed techniques. As a consequence, the benchmarking of the effectiveness of the proposed techniques is not straightforward. In spite of lack of comparative experimental results, we observe that the post processing of i-vectors with duration-matched development data gives considerable gain in verification accuracy over baseline [22]. Moreover, this approach does not add additional computational overhead and provides a simple solution suitable for real-world application.

6 Future Research Directions

The state-of-the-art techniques in ASV research are primarily suitable for a large amount of speech in training and test. ASV techniques, which are best suited with a sufficiently large amount of data, may be inappropriate for a smaller amount of speech. Future investigations should be carried out considering the general limitations of short utterances. Here we highlight some possible research directions for solving this very important and challenging real-world problem.

- **Deep Neural Network (DNN)** based approaches can be employed for discovering features. There are successful efforts in recent past for such ASV systems [106, 107]. In contemporary literatures, DNNs have been incorporated in different stages of ASV system. A speech-based DNN is used to extract bottleneck (BN) features from an inner layer restricted in dimensionality [108–112]. The estimation of BW statistics from traditional GMM is replaced by a DNN performing resembling tasks [35]. BN features were first proposed for large vocabulary speech recognition in conjunction with conventional acoustic features, such as MFCC [113], and have more recently shown similar improvement in language and speaker identification applications [108, 111, 112]. DNN-based speaker modeling architectures successfully attempted to accommodate phonetic granularity at different levels [107]. Recently, DNN-based end-to-end system has been explored for text-dependent speaker recognition with a fixed pass-phrase [114]. These successful application of DNNs in speaker verification encourage further study in this field for further advancements in short utterance speaker verification.
- **Metric Learning Technique** The present state-of-the-art ASV systems uses log likelihood of GPLDA projected i-vectors to calculate the ASV scores [80, 90]. The reports indicates that the performance for ASV can be improved considerably with better projection of speaker-model vectors and score calculation. Although few attempts on specialized speaker-model vector projection for short duration segments has been done [22, 90, 95, 98, 100], but comparatively less effort has been made in improving score computation [93].

  Different distance or similarity metrics can be explored to measure the score of i-vector representation of speech utterances [115]. The idea here is to find a distance metric function which minimizes the distances between i-vectors of the same speakers and maximizes the distances between different speakers. The development data with short speech segments can be used to train the metric and subsequently, the learned metric can be used for measuring the similarity of i-vectors in a pairwise manner. Previous studies in distance metric method in speaker verification do not focus on short utterance

problem [116, 117], however, the approach could be explored further to investigate the similarity between i-vectors of small speech segments.

- **Sparse methods** The limited data conditions in speaker verification leads to sparsity in sufficient statistics estimation which is successively used in i-vector estimation [49, 98]. Methods developed to handle the sparsity issue such as dictionary learning [118], sparse representation [119] can be investigated to effectively process and represent the speech data for short utterances. Previously, these techniques were successfully applied for the classification of partial data (e.g., partial face image), and for this reason, sparse methods are potentially useful for short utterance speaker verification.
- **Dimensionality reduction techniques** In speaker verification with i-vectors, LDA or variant of PLDA is used for reducing the dimension of i-vectors. This process, by nature, removes redundant information, for example, session variability effects. The research in dimensionality reduction studies is much more extensive than its application in ASV context [120]. Recently, generalized discriminant analysis is investigated as an alternative of standard LDA [121]. The dimensionality reduction process requires further investigation as the currently used techniques are not necessarily optimal for i-vectors, estimated from short utterances with larger uncertainty.
- **Miscellaneous Opportunities** Conventional short-term spectral features capture time-localized snapshot of speech production system which is highly dependent on the spoken text or phonetic information. To address the short duration issue, we need a feature extraction approach invariant to the underlying phonetic content. For example, investigation of speaker-specific information with less dependency on the acoustic class or spoken content could be useful in this context.

ASV performance with short segments in presence of noise is an important concern for the deployment of ASV systems in real-life situations. So far, only a few studies are available with short utterance speaker verification in noisy conditions which opens up scopes for future research in this very relevant problem [122, 123].

Most of the research in speaker verification dwell on the ASV performance on a given speech corpus consisting data for train and test. However, less study has been conducted to investigate whether a given speech segment is adequate for enrollment or test. Investigation on the reliability of speech segments for confidence measure in ASV task should be carried out to assist the decision making process in recognition systems by providing an objective quality of the speech data under consideration. Standardization of the amount of training and verification data for achieving satisfactory ASV performance remains another challenging problem in speaker verification task.

7 CONCLUSION

The research effort to tackle the problem of short utterance for speaker verification context has been significantly increased in recent years. The problem is addressed in different subsystem levels of ASV framework like feature extraction, speaker modeling, score normalization, score calibration, etc. In this paper, we provide a detailed literature review of the related research work emphasizing the recent studies. This extensive review can help the interested researchers to identify the key challenges of the duration variability in ASV systems and to choose the efficient methods, specific to a subsystem level development. Moreover, this review will be useful in promoting further research in several potentially interesting directions.

Acknowledgement

This work is partially supported by Indian Space Research Organization (ISRO), Government Of India.

8 References

1 Kinnunen, T., Li, H.: 'An overview of text-independent speaker recognition: From features to supervectors', *Speech Communication*, 2010, **52**, (1), pp. 12–40
2 Campbell.Jr, J.P.: 'Speaker recognition: A tutorial', *Proceedings of the IEEE*, 1997, **85**, (9), pp. 1437–1462
3 Chakroborty, S., Saha, G.: 'Improved text-independent speaker identification using fused MFCC & IMFCC feature sets based on Gaussian filter', *International Journal of Signal Processing*, 2009, **5**, (1), pp. 11–19
4 Hébert, M. 'Text-dependent speaker recognition'. (Springer, 2008. pp. 743–762
5 Shetty, M.: 'ICICI bank to roll out voice authentication', *available at: wwwtimesofindiaindiatimescom/business/india-business/ICICI-Bank-to-roll-out-voice-authentication/articleshow/46818823cms*, April 6, 2015 [Online],
6 Loshin, P.: 'Barclays replaces passwords with voice authentication', *available at: http://searchsecuritytechtargetcom/news/450301866/Barclays-replaces-passwords-with-voice-authentication*, 3 Aug 2016 [Online],
7 Mandasari, M.I., McLaren, M., van Leeuwen, D.A.: 'Evaluation of i-vector speaker recognition systems for forensic application.', *Proc INTERSPEECH*, 2011, pp. 21–24
8 Larcher, A., Lee, K.A., Ma, B., Li, H.: 'RSR2015: Database for text-dependent speaker verification using multiple pass-phrases.', *Proc INTERSPEECH*, 2012, pp. 1580–1583
9 Jayanna, H., Prasanna, S.M.: 'Analysis, feature extraction, modeling and testing techniques for speaker recognition', *IETE Technical Review*, 2009, **26**, (3), pp. 181–190
10 Hansen, J.H.L., Hasan, T.: 'Speaker recognition by Machines and Humans: A tutorial review', *IEEE Signal Processing Magazine*, 2015, **32**, (6), pp. 74–99
11 Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: 'Speaker verification using adapted Gaussian mixture models', *Digital Signal Processing*, 2000, **10**, (1), pp. 19–41
12 Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P.: 'Front-end factor analysis for speaker verification', *IEEE Transactions on Audio, Speech, and Language Processing*, 2011, **19**, (4), pp. 788–798
13 Satyanarayana, P.: 'Short segment analysis of speech for enhancement', *PhD dissertation, Indian Institute of Technology Madras, India*, Feb. 1999,
14 Yegnanarayana, B., Prasanna, S.M., Zachariah, J.M., Gupta, C.S.: 'Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system', *IEEE Transactions on Speech and Audio Processing*, 2005, **13**, (4), pp. 575–582
15 Prasanna, S.M., Gupta, C.S., Yegnanarayana, B.: 'Extraction of speaker-specific excitation information from linear prediction residual of speech', *Speech Communication*, 2006, **48**, (10), pp. 1243–1261
16 Murty, K.S.R., Yegnanarayana, B.: 'Combining evidence from residual phase and mfcc features for speaker recognition', *IEEE Signal Processing Letters*, 2006, **13**, (1), pp. 52–55
17 Rabiner, L.R., Juang, B.H.: 'Fundamentals of speech recognition'. (PTR Prentice Hall, 1993)
18 Reynolds, D.A.: 'Experimental evaluation of features for robust speaker identification', *IEEE Transactions on Speech and Audio Processing*, 1994, **2**, (4), pp. 639–643
19 Sahidullah, M., Saha, G.: 'Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition', *Speech Communication*, 2012, **54**, (4), pp. 543–565
20 Sahidullah, M., Saha, G.: 'A novel windowing technique for efficient computation of MFCC for speaker recognition', *IEEE Signal Processing Letters*, 2013, **20**, (2), pp. 149–152
21 Mandasari, M.I., Saeidi, R., McLaren, M., van Leeuwen, D.: 'Quality measure functions for calibration of speaker recognition systems in various duration conditions', *IEEE Transactions on Audio, Speech, and Language Processing,*, 2013, **21**, (11), pp. 2425–2438
22 Kanagasundaram, A., Dean, D., Sridharan, S., Gonzalez.Dominguez, J., Gonzalez.Rodriguez, J., Ramos, D.: 'Improving short utterance i-vector speaker verification using utterance variance modelling and compensation techniques', *Speech Communication*, 2014, **59**, pp. 69–82
23 Sahidullah, M., Kinnunen, T.: 'Local spectral variability features for speaker verification', *Digital Signal Processing*, 2016, **50**, pp. 1–11
24 Atal, B.S.: 'Automatic speaker recognition based on pitch contours', *The Journal of the Acoustical Society of America*, 1972, **52**, (6B), pp. 1687–1697
25 Mary, L., Rao, K., Gangashetty, S., Yegnanarayana, B.: 'Neural network models for capturing duration and intonation knowledge for language and speaker identification', *Proc ICCNS*, 2004,
26 Farahani, F., Georgiou, P.G., Narayanan, S.S.: 'Speaker identification using supra-segmental pitch pattern dynamics', *Proc ICASSP*, 2004, **1**, pp. I–89
27 Davis, S.B., Mermelstein, P.: 'Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences', *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1980, **28**, (4), pp. 357–366
28 Hermansky, H.: 'Perceptual linear predictive (PLP) analysis of speech', *The Journal of the Acoustical Society of America*, 1990, **87**, (4), pp. 1738–1752
29 Dişken, G., Tüfekçi, Z., Saribulut, L., Çevik, U.: 'A review on feature extraction for speaker recognition under degraded conditions', *IETE Technical Review*, 2016, pp. 1–12
30 Furui, S.: 'Cepstral analysis technique for automatic speaker verification', *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 1981, **29**, (2), pp. 254–272
31 Reynolds, D.A., Rose, R.C.: 'Robust text-independent speaker identification using gaussian mixture speaker models', *IEEE transactions on speech and audio processing*, 1995, **3**, (1), pp. 72–83
32 Kenny, P., Ouellet, P., Dehak, N., Gupta, V., Dumouchel, P.: 'A study of interspeaker variability in speaker verification', *IEEE Transactions on Audio, Speech, and Language Processing*, 2008, **16**, (5), pp. 980–988

33 Kenny, P.: 'Bayesian speaker verification with heavy-tailed priors.', *Proc Odyssey*, 2010, p. 14
34 Deng, L., Yu, D.: 'Deep learning: methods and applications', *Foundations and Trends in Signal Processing*, 2014, **7**, (3–4), pp. 197–387
35 Kenny, P., Gupta, V., Stafylakis, T., Ouellet, P., Alam, J.: 'Deep neural networks for extracting Baum-Welch statistics for speaker recognition', *Proc Odyssey*, 2014, pp. 293–298
36 Lei, Y., Scheffer, N., Ferrer, L., McLaren, M.: 'A novel scheme for speaker recognition using a phonetically-aware deep neural network', *Proc ICASSP*, 2014, pp. 1695–1699
37 Tirumala, S.S., Shahamiri, S.R.: 'A review on deep learning approaches in speaker identification', *Proceedings of the 8th International Conference on Signal Processing Systems*, 2016, pp. 142–147
38 Variani, E., Lei, X., McDermott, E., Moreno, I.L., Gonzalez.Dominguez, J.: 'Deep neural networks for small footprint text-dependent speaker verification', *Proc ICASSP*, 2014, pp. 4052–4056
39 Li, L., Wang, D., Zhang, Z., Zheng, T.F.: 'Deep speaker vectors for semi text-independent speaker verification', *preprint arXiv:150506427*, 2015,
40 Poddar, A., Sahidullah, M., Saha, G.: 'Performance comparison of speaker recognition systems in presence of duration variability.', *Proc IEEE INDICON*, 2015, pp. 1–6
41 Sizov, A., Lee, K.A., Kinnunen, T.: 'Unifying probabilistic linear discriminant analysis variants in biometric authentication', *Proc S+SSPR*, 2014, pp. 464–475
42 Hasan, T., Saeidi, R., Hansen, J.H., van Leeuwen, D.: 'Duration mismatch compensation for i-vector based speaker recognition systems', *Proc ICASSP*, 2013, pp. 7663–7667
43 NIST: 'The NIST year 2010 speaker recognition evaluation plan', *techrep, NIST*, 2010,
44 NIST: 'The NIST year 2008 speaker recognition evaluation plan', *techrep, NIST*, 2008,
45 Kanagasundaram, A., Vogt, R., Dean, D.B., Sridharan, S., Mason, M.W.: 'I-vector based speaker recognition on short utterances', *Proc INTERSPEECH*, 2011, pp. 2341–2344
46 Hautamäki, V., Cheng, Y.C., Rajan, P., Lee, C.H.: 'Minimax i-vector extractor for short duration speaker verification.', *Proc INTERSPEECH*, 2013, pp. 3708–3712
47 Poorjam, A.H., Saeidi, R., Kinnunen, T., Hautamäki, V.: 'Incorporating uncertainty as a quality measure in i-vector based language recognition', *Proc Odyssey*, 2016, pp. 74–80
48 Shum, S. 'Unsupervised methods for speaker diarization'. Massachusetts Institute of Technology, 2011
49 Poddar, A., Sahidullah, M., Saha, G.: 'An adaptive i-vector extraction for speaker verification with short utterance', *International Conference on Pattern Recognition and Machine Intelligence*, 2017, **7**, pp. 1–6
50 Schwarz, P., Matejka, P., Cernocky, J.: 'Hierarchical structures of neural networks for phoneme recognition', *in Proc ICASSP*, 2006, **1**, pp. I–I
51 Sarkar, A.K., Matrouf, D., Bousquet, P.M., Bonastre, J.F.: 'Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification.', *Proc INTERSPEECH*, 2012,
52 Chen, Y., Tang, Z.M.: 'The speaker recognition of noisy short utterance', *ICISBDE*, 2013, pp. 666–671
53 Li, Z.Y., Zhang, W.Q., Liu, J.: 'Multi-resolution time frequency feature and complementary combination for short utterance speaker recognition', *Multimedia Tools and Applications*, 2015, **74**, (3), pp. 937–953
54 Alam, M.J., Kenny, P., Stafylakis, T.: 'Combining amplitude and phase-based features for speaker verification with short duration utterances.', *Proc INTERSPEECH*, 2015, pp. 249–253
55 Das, R.K., Abhiram, S., Prasanna, S.M., Ramakrishnan, A.: 'Combining source and system information for limited data speaker verification.', *Proc INTERSPEECH*, 2014, pp. 1836–1840
56 Das, R.K., Pati, D., Prasanna, S.M.: 'Different aspects of source information for limited data speaker verification', *Proc NCC*, 2015, pp. 1–6
57 Das, R.K., Mahadeva.Prasanna, S.: 'Exploring different attributes of source information for speaker verification with limited test data', *The Journal of the Acoustical Society of America*, 2016, **140**, (1), pp. 184–190
58 Rosenberg, A.E., Soong, F.K.: 'Evaluation of a vector quantization talker recognition system in text independent and text dependent modes', *Computer Speech & Language*, 1987, **2**, (3-4), pp. 143–157
59 Hautamki, V., Kinnunen, T., Krkkinen, I., Saastamoinen, J., Tuononen, M., Frnti, P.: 'Maximum a posteriori adaptation of the centroid model for speaker verification', *IEEE Signal Processing Letters*, 2008, **15**, pp. 162–165
60 Kinnunen, T., Karpov, E., Franti, P.: 'Real-time speaker identification and verification', *IEEE Transactions on Audio, Speech, and Language Processing*, 2006, **14**, (1), pp. 277–288
61 Li, K., Wrench, E.: 'An approach to text-independent speaker recognition with short utterances', *Proc ICASSP*, 1983, **8**, pp. 555–558
62 Wagner, M.: 'Combined speech recognition/speaker-verification system with modest training requirements', *Proc SST*, 1996, pp. 139–143
63 Matsui, T., Furui, S.: 'Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMM's', *IEEE Transactions on Audio, Speech, and Language Processing*, 1994, **2**, (3), pp. 456–459
64 Tran, D., Wagner, M., Van.Le, T.: 'A proposed decision rule for speaker recognition based on fuzzy c-means clustering.', *Proc ICSLP*, 1998,
65 Li, K.P., Porter, J.E.: 'Normalizations and selection of speech segments for speaker recognition scoring', *Proc ICASSP*, 1988, pp. 595–598
66 Merlin, T., Bonastre, J.F., Fredouille, C.: 'Non directly acoustic process for costless speaker recognition and indexation', *Proc IWICTA*, 1999, **29**
67 Ferrer, L., et al.: 'Modeling duration patterns for speaker recognition', *Proc Eurospeech*, 2003, pp. 2017–2020
68 Larcher, A., Bonastre, J.F., Mason, J.S.: 'Short utterance-based video aided speaker recognition', *Proc WMSP*, 2008, pp. 897–901
69 Vogt, R.J., Lustri, C.J., Sridharan, S.: 'Factor analysis modeling for speaker verification with short utterances', *Proc Odyssey*, 2008,
70 Vogt, R., Sridharan, S.: 'Minimising speaker verification utterance length through confidence based early verification decisions', *Proc ICB*, 2009, pp. 454–463
71 Vogt, R., Sridharan, S., Mason, M.: 'Making confident speaker verification decisions with minimal speech', *IEEE Transactions on Audio, Speech, and Language Processing*, 2010, **18**, (6), pp. 1182–1192
72 Stadelmann, T., Freisleben, B.: 'Dimension-decoupled Gaussian mixture model for short utterance speaker recognition', *in Proc ICPR*, 2010, pp. 1602–1605
73 Fauve, B.G., Evans, N.W., Pearson, N., Bonastre, J.F., Mason, J.S.: 'Influence of task duration in text-independent speaker verification.', *Proc INTERSPEECH*, 2007, pp. 794–797
74 Fauve, B.G., Evans, N.W., Mason, J.S.: 'Improving the performance of text-independent short duration SVM-and GMM-based speaker verification.', *Proc Odyssey*, 2008, p. 18
75 McLaren, M., Vogt, R., Baker, B., Sridharan, S., Sridharan, S.: 'Experiments in SVM-based speaker verification using short utterances.', *Proc Odyssey*, 2010, p. 17
76 Krishnamoorthy, P., Jayanna, H.S., Prasanna, S.R.M.: 'Speaker recognition under limited data condition by noise addition', *Expert Systems with Applications*, 2011, **38**, (10), pp. 13487–13490
77 Zhang, C., Zheng, T.F.: 'A Fishervoice based feature fusion method for short utterance speaker recognition', *Proc ChinaSIP*, 2013, pp. 165–169
78 Larcher, A., Bonastre, J.F., Mason, J.S.: 'Constrained temporal structure for text-dependent speaker verification', *Digital Signal Processing*, 2013, **23**, (6), pp. 1910–1917
79 Zhang, W.Q., Zhao, J., Zhang, W.L., Liu, J.: 'Multi-scale kernels for short utterance speaker recognition', *in Proc ISCSLP*, 2014, pp. 414–417
80 Kanagasundaram, A., Vogt, R.J., Dean, D.B., Sridharan, S.: 'PLDA based speaker recognition on short utterances', *Proc Odyssey*, 2012,
81 Larcher, A., Bousquet, P.M., Lee, K.A., Matrouf, D., Li, H., Bonastre, J.F.: 'I-vectors in the context of phonetically-constrained short utterances for speaker verification', *Proc ICASSP*, 2012, pp. 4773–4776
82 Fatima, N., Zheng, T.F.: 'Vowel-category based short utterance speaker recognition', *Proc ICSAI*, 2012, pp. 1774–1778
83 Fatima, N., Zheng, T.F.: 'Syllable category based short utterance speaker recognition', *Proc ICALIP*, 2012, pp. 436–441
84 Larcher, A., Lee, K.A., Ma, B., Li, H.: 'Phonetically-constrained PLDA modeling for text-dependent speaker verification with multiple short utterances', *Proc ICASSP*, 2013, pp. 7673–7677
85 Bharathi, B., Nagarajan, T.: 'GMM and i-vector based speaker verification using speaker-specific-text for short utterances', *Proc TENCON*, 2013, pp. 1–4
86 Domínguez, J.G., Zazo, R., González.Rodríguez, J. 'On the use of total variability and probabilistic linear discriminant analysis for speaker verification on short utterances'. (Springer, 2012. pp. 11–19
87 Kanagasundaram, A., Dean, D., González.Domínguez, J., Sridharan, S., Ramos, D., Gonzalez.Rodriguez, J.: 'Improving short utterance based i-vector speaker recognition using source and utterance-duration normalization techniques', *Proc INTERSPEECH*, 2013,
88 Huber, P.J.: 'A robust version of the probability ratio test', *The Annals of Mathematical Statistics*, 1965, **36**, (6), pp. 1753–1758
89 Soldi, G., Bozonnet, S., Alegre, F., Beaugeant, C., Evans, N.: 'Short-duration speaker modelling with phone adaptive training', *Proc Odyssey*, 2014,
90 Kanagasundaram, A., Dean, D., Sridharan, S.: 'Improving PLDA speaker verification with limited development data', *Proc ICASSP*, 2014, pp. 1665–1669
91 Biagetti, G., Crippa, P., Curzi, A., Orcioni, S., Turchetti, C.: 'Speaker identification with short sequences of speech frames', *Proc ICPRAM*, 2015, pp. 178–185
92 Hong, Q., Li, L., Li, M., Huang, L., Wan, L., Zhang, J.: 'Modified-prior PLDA and score calibration for duration mismatch compensation in speaker recognition system.', *Proc INTERSPEECH*, 2015, pp. 1037–1041
93 Mandasari, M.I., Saeidi, R., van Leeuwen, D.A.: 'Quality measures based calibration with duration and noise dependency for speaker recognition', *Speech Communication*, 2015, **72**, pp. 126–137
94 Das, R.K., Jelil, S., Prasanna, S.M.: 'Significance of constraining text in limited data text-independent speaker verification', *Proc SPCOM*, 2016, pp. 1–5
95 Mamodiya, S., Kumar, L., Das, R.K., Prasanna, S.M.: 'Exploring acoustic factor analysis for limited test data speaker verification', *Proc TENCON*, 2016, pp. 1397–1401
96 Manam, A.B., Revanth, T.S., Das, R.K., Prasanna, S.M.: 'Speaker verification using acoustic factor analysis with phonetic content compensation in limited and degraded test conditions', *Proc TENCON*, 2016, pp. 1402–1406
97 Li, L., Wang, D., Zhang, C., Zheng, T.F.: 'Improving short utterance speaker recognition by modeling speech unit classes', *IEEE Transactions on Audio, Speech, and Language Processing*, 2016, **24**, (6), pp. 1129–1139
98 Li, W., Fu, T., You, H., Zhu, J., Chen, N.: 'Feature sparsity analysis for i-vector based speaker verification', *Speech Communication*, 2016, **80**, pp. 60–70
99 Ando, A., Asami, T., Yamaguchi, Y., Aono, Y.: 'Speaker recognition in duration-mismatched condition using bootstrapped i-vectors', *Proc APSIPA*, 2016, pp. 1–4
100 Kanagasundaram, A., Dean, D., Sridharan, S., Ghaemmaghami, H., Fookes, C.: 'A study on the effects of using short utterance length development data in the design of GPLDA speaker verification systems', *International Journal of Speech Technology*, 2017, pp. 1–13
101 Ma, J., Sethu, V., Ambikairajah, E., Lee, K.A.: 'Duration compensation of i-vectors for short duration speaker verification', *Electronics Letters*, 2017, **53**, (6), pp. 405–407
102 Yang, I.H., Heo, H.S., Yoon, S.H., Yu, H.J.: 'Applying compensation techniques on i-vectors extracted from short-test utterances for speaker verification using

deep neural network', *in Proc ICASSP*, 2017, pp. 5490–5494
103 Hasan, T., Hansen, J.H.: 'Acoustic factor analysis for robust speaker verification', *IEEE Transactions on Audio, Speech, and Language Processing*, 2013, **21**, (4), pp. 842–853
104 Li, L., Wang, D., Zhang, X., Zheng, T.F., Jin, P.: 'System combination for short utterance speaker recognition', *arXiv preprint arXiv:160309460*, 2016,
105 Villalba, J., Ortega, A., Miguel, A., Lleida, E.: 'Analysis of speech quality measures for the task of estimating the reliability of speaker verification decisions', *Speech Communication*, 2016, **78**, pp. 42–61
106 Kanagasundaram, A., Dean, D., Sridharan, S., Fookes, C.: 'Dnn based speaker recognition on short utterances', *preprint arXiv:161003190*, 2016,
107 Tian, Y., He, L., Cai, M., Zhang, W.Q., Liu, J.: 'Deep neural networks based speaker modeling at different levels of phonetic granularity', *in Proc ICASSP*, 2017, pp. 5440–5444
108 McLaren, M., Lei, Y., Ferrer, L.: 'Advances in deep neural network approaches to speaker recognition', *Proc ICASSP*, 2015, pp. 4814–4818
109 Snyder, D., Garcia.Romero, D., Povey, D.: 'Time delay deep neural network-based universal background models for speaker recognition', *Proc ASRU*, 2015, pp. 92–97
110 Garcia.Romero, D., Zhang, X., McCree, A., Povey, D.: 'Improving speaker recognition performance in the domain adaptation challenge using deep neural networks', *Proc SLT*, 2014, pp. 378–383
111 Richardson, F., Reynolds, D., Dehak, N.: 'Deep neural network approaches to speaker and language recognition', *IEEE Signal Processing Letters*, 2015, **22**, (10), pp. 1671–1675
112 Richardson, F., Reynolds, D., Dehak, N.: 'A unified deep neural network for speaker and language recognition', *preprint arXiv:150400923*, 2015,
113 Hinton, G., et al.: 'Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups', *IEEE Signal Processing Magazine*, 2012, **29**, (6), pp. 82–97
114 Heigold, G., Moreno, I., Bengio, S., Shazeer, N.: 'End-to-end text-dependent speaker verification', *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, 2016, pp. 5115–5119
115 Yang, L., Jin, R. 'Distance metric learning: A comprehensive survey'. (Michigan State University, 2006.
116 Li, L., Wang, D., Xing, C., Zheng, T.F.: 'Max-margin metric learning for speaker recognition', *Chinese Spoken Language Processing (ISCSLP), 2016 10th International Symposium on*, 2016, pp. 1–4
117 Lei, Z., Wan, Y., Luo, J., Yang, Y.: 'Mahalanobis metric scoring learned from weighted pairwise constraints in i-vector speaker recognition system.', *INTERSPEECH*, 2016, pp. 1815–1819
118 Tosic, I., Frossard, P.: 'Dictionary learning', *IEEE Signal Processing Magazine*, 2011, **28**, (2), pp. 27–38
119 Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: 'Robust face recognition via sparse representation', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, **31**, (2), pp. 210–227
120 Van.Der.Maaten, L., Postma, E., Van den Herik, J.: 'Dimensionality reduction: a comparative', *Journal of Machine Learning Research*, 2009, **10**, pp. 66–71
121 Bahmaninezhad, F., Hansen, J.H.L.: 'Generalized discriminant analysis (GDA) for improved i-vector based speaker recognition', *INTERSPEECH*, 2016, pp. 3643–3647
122 Aldhaheri, R.W., Al.Saadi, F.E.: 'Robust text-independent speaker recognition with short utterance in noisy environment using SVD as a matching measure', *Journal of King Saud University-Computer and Information Sciences*, 2004, **17**, pp. 25–44
123 Chen, Y., Tang, Z.: 'Speaker recognition of noisy short utterance based on speech frame quality discrimination and three-stage classification model', *International Journal of Control and Automation*, 2015, **8**, (3), pp. 135–146