An Overview of Robustness Related Issues in Speaker Recognition

Thomas Fang Zheng[*1], Qin Jin[2], Lantian Li[1], Jun Wang[1], and Fanhu Bie[1]
[1] Center for Speech and Language Technologies
Division of Technical Innovation and Development, Tsinghua National Laboratory for Information Science and Technology;
Research Institute of Information Technology; Department of Computer Science and Technology
Tsinghua University, Beijing, 100084, China
[*] Corresponding author e-mail: fzheng@tsinghua.edu.cn
[2] Multimedia Computing Lab, School of Information, Renmin University of China
E-mail: qjin@ruc.edu.cn

***Abstract*— Speaker recognition technologies have been improved rapidly in recent years. However, critical robustness issues need to be addressed when they are applied in practical situations. This paper provides an overview of technologies dealing with robustness related issues in automatic speaker recognition. We first categorize the robustness issues into three categories, including environment-related, speaker-related and application-oriented issues. For each category, we then describe the current hot topics, existing technologies, and potential research focuses in the future.**

I. INTRODUCTION

Spoken language is the most natural way we humans communicate with each other. There is rich information conveyed in spoken language, including language information (linguistic contents etc.), speaker information (identity, emotion, physiological characteristics etc.), environmental information (background, channel etc.) and so on. We humans can effortlessly decode most of such information although such rich information is encoded in a complex form. This human ability has inspired a lot of research work to automatically extract and process the richness of information in spoken language. Automatic speaker recognition, the process of recognizing a person's identity from his or her voice, has been one of the most active research areas in spoken language processing [1, 2]. Automatic speaker recognition technologies have wide applications such as access control, transaction authentication, voice-based information retrieval, recognition of perpetrator in forensic analysis, and personalization of user devices etc. Speaker diarization [3] attempts to find speakers turn takings in a conversation. It is an extension of the "classical" speaker recognition technologies applied in multiparty conversations.

There have been several very useful survey or tutorial papers for speaker recognition in the past, including Furui's overview [1] which presents the principles of speaker recognition and typical approaches for text-dependent and text-independent speaker recognition, Campbell's tutorial [2] presents in-detail descriptions about designing and building an automatic speaker recognition systems and the most recent overview by Kinnunen and Li [4] which gives a thorough overview of text-independent speaker recognition technologies covering speaker features and speaker models. In this paper we present an overview of speaker recognition technologies with an emphasis on dealing with robustness issues. We first categorize the robustness issues into three categories, including environment-related issues, speaker-related issues and application-oriented issues. For each category, we then describe the current hot topics, existing technologies, and potential research focuses in the future.

II. ENVIRONMENT-RELATED ROBUSTNESS ISSUES

Many factors can influence the performance of speaker verification systems. Environmental noise and channel mismatch [5 - 7] are the two most common factors, especially in many real applications. On the one hand, speech audio recorded from real environments often contains different types of environmental noise such as background white noise, music, or interfering speech etc. Environmental noise has adverse impact on speaker modeling. On the other hand, speech audio files are often recorded through various transmission channels using different types of microphones. Channel mismatch will greatly influence the character of speech signal. In the Speaker Recognition Evaluations organized by NIST [8], the channel mismatch issue is always regarded as one of the challenges in evaluations. To encourage the research dealing with channel mismatch, different recording equipment and transmission channel have been utilized in collecting the evaluation data [9, 10]. In real applications, the prior knowledge of environmental noise and transmission channel is not available in advance, which makes pre-training noise/channel models impossible.

A. Noise Robustness

Recording or environmental noise degrades speaker verification performance. The research on noise robustness can be categorized into two directions: improving the feature robustness and model robustness against noise. Spectral Subtraction [11, 12] is effective in dealing with stable noises, but fails when the noise is not stable. RASTA filtering [13] uses band-pass filtering in the log spectral domain to remove

slow channel variation noises. There are also algorithms such as PCA (Principle Component Analysis) [14], LDA (Linear Discriminant Analysis) [15] and HLDA (Heteroscedastic Linear Discriminant Analysis) [16] applied in feature domain to improve robustness. The research on model robustness against noise usually adopts model compensation [17, 18] algorithms to decrease the mismatch between test sets and training sets.

B. Channel Mismatch

Channel mismatch is another salient factor that influences the recognition robustness. In real applications, speech utterances are often recorded with various types of microphones (desktop microphone or head phone, etc.), and these speech signals are changed in some degree due to the different transmission channels. Research dealing with the channel mismatch of speaker verification tasks can be categorized into three directions: feature transformation [7, 19, 20], model compensation [21, 22] and score normalization [23, 24]. CMS (Cepstral Mean Subtraction) [6] or Cepstral Mean Normalization (CMN) which subtracts the mean value of each feature vector over the entire utterance is the most simple and common method used in many speaker verification systems. The channel variations are considered to be stable over the entire utterance in these methods. Feature mapping [7] which maps the features to a channel-independent feature space and feature warping which modifies the short-term feature distribution to follow a reference distribution [25] are also effective methods but with more complex implementation. SMS (Speaker Model Synthesis) is popular in GMM-UBM systems [26], which transforms models from one channel to another according to UBM deviations between channels. [22] proposed a modified SMS method based on a cohort dataset.

JFA (Joint Factor Analysis) [26], a more comprehensive statistical approach, has gained much success in speaker verification. The speaker variations and channel (session) variations are modeled as independent variables spanning in a low-rank subspace, which defines the speaker- and channel-variations as two independent random variables following *a priori* standard Gaussian distributions. Then the factors are inferred the posterior probability of the speaker- and channel-variations from the given speech. [27] assumed channel factors in a low-rank space, leaving the speaker factors in a full-rank space. [28] proposed a simple implementation for the subspace inferring.

The i-vector method [29] considers that the speaker and channel variations cannot be separated by JFA (channel variations also contains speaker information). So in i-vector methodology, a low-rank total variability space is defined to represent speaker- and channel-variations at the same time, and the speaker utterance is represented by an i-vector which is derived by inferring the posterior distribution of the total variance factor. There is no distinction between speaker effects and channel effects in GMM supervector space. Both speaker- and channel-variations are retained in i-vector. However, the total representation will lead to less discrimination among speakers due to channel variations. Therefore, many inter-channel compensation methods especially some popular discriminative approaches are employed to extract more accentuate speaker information. WCCN (With-in Class Covariance Normalization) [30] and LDA (Linear Discriminant Analysis) [29, 33] are both linear transform to optimize the linear kernels. NAP (Nuisance Attribute Projection) [27] is to find the projection optimized by minimizing the difference among channels. The most recent research focuses on the PLDA (Probabilistic Linear Discriminant Analysis) [31, 32], which improves the performance of i-vector system greatly. PLDA is a probabilistic version of LDA, and also is a generative model that utilizes a prior distribution on the speaker- and channel-variations. Although PLDA has achieved great success in speaker verification, the limitation is that PLDA still takes a Gaussian distribution as the prior factor that may not always be true in reality.

Channel mismatch was also studied within neural network and SVM. [34] proposed to reduce channel impact for verification systems based on neural networks by eliminating a proportion of hidden nodes. Neural network framework will be a very prospective direction in speaker verification research.

III. Speaker-Related Robustness Issues

Speaker-related variability, such as gender, physical state (cold or laryngitis), speaking style (emotion, speaking rate, etc.), cross-language, accent and session variations, is one of the main difficulties in speech signal processing. How they correlate with each other and what are the key factors in speech realization are real concerns in speech research [35]. The current mainstream research can be divided into five directions (gender, physical condition, speaking style, cross-lingual, aging etc.) as described in the following subsections

A. Gender

Speaker recognition systems attain their best accuracy when trained with gender dependent (GD) features and tested with known gender trails. However, in real applications, gender labels are often not given. Therefore, how to design a system that does not make use of the gender labels both in training and test has highly practical significance. [36] addressed the problem of designing a fully gender independent (GI) speaker recognition system. It relies on discriminative training, where the trails are i-vector pairs, and the discrimination is between the hypothesis that the pair of feature vectors in the trial belong to the same speaker or different speakers. They demonstrated that this pairwise discriminative training can be interpreted as a procedure that estimates the parameters of the best (second order) approximation of the log-likelihood ratio score function, and that a pairwise SVM can be used for training a GI system. The results show that the performance of this fully GI system is slightly worse than that of a fully GD systems. A novel GI PLDA classifier for i-vector based speaker recognition was proposed and evaluated in [37]. This approach uses the source-normalization (SN) for variation that separates genders as a pre-processing step for i-vector based PLDA

classification. Experimental results on the NIST 2010 SRE dataset have demonstrated that it can reach comparable performance compared with a typical GD configuration.

B. Physical Conditions

Speech is a behavioral signal that may not be consistently reproduced by a speaker and can be affected by a speaker's physical conditions. This variability aspect may be due to illness such as cold, nasal congestion, laryngitis and other behavioral variations.

As early as 1996, Tull with his team has carried out the related researches on the "cold-affected" speech in speaker recognition [38, 39]. They analyzed the differences between "cold-affected" speech and normal/healthy speech from the resonances of the vocal tract, fundamental frequency (measurements of voice pitch), phonetic differences and mel-cepstral coefficients. They found that "cold-speech" shows some noisy portions in the acoustic signal that are not present in the normal/healthy signals. These noisy portions are caused by hoarseness and coughing. Besides, they also analyzed phonetic contrasts and looked at differences in formant patterns. Phonetic transcriptions of cold and normal sessions reflect changes in place of articulation. Perceptual and acoustic analyses revealed that pauses and epenthetic syllables are not constant throughout all sessions. These differences suggested that the cold introduces another level of intra-speaker variability that needs to be addressed in the design of speaker recognition systems.

Nowadays, however, research in this direction is still rare and most researches are just focused on feature level [40, 41]. The main reason is that this type of speech database is difficult to collect and organize. But in real life situations, a person's physical condition is variable and will have great effect on the performance of speaker recognition systems. Therefore, research on robustness to the speaker-physical-condition variability has very practical importance.

C. Speaking Style

Speaking style usually comes from speaker's spontaneous or no-prompted speech and contains abundant speaker individual information. In the field of speech recognition, there has been a lot of research on the speaking style [42 - 44]. The main method is to choose more emotion-related features and train multi-style speaking models. In recent years, many researchers attempted to apply these methods in speaker recognition systems. In this section, we will review the speaking style analysis of speaker recognition from three aspects:

a) Emotion

Emotion is an intrinsic nature of human beings and changes in the rendering forms of speech signals significantly. Compared with other intra-speaker variations such as the speaking rate and speech tones, emotion tends to cause more substantial variation on speech properties, such as harmonic forms, formant structures and the entire temporal-spectral patterns [45].

A multitude of research have been conducted to address the emotion variations. The first category involves analysis of various emotion-related acoustic factors such as prosody and voice quality [46], pitch [47, 48], duration and sound intensity [48]. The second category involves various emotion-compensation methods for models and scores. An early investigation was supported by the European VERIVOX project [49, 50], where the researchers proposed a 'structured training' that elicits enrollment speech in various speaking style. By training the speaker models with the elicited multi-emotional speech, the authors reported reasonable performance improvements. This method, however, has poor-interactivity and is unacceptable in practice. Wu *et al*. [47] presented an emotion-added model training method, where a few amount of emotional data were used to train emotion-dependent models. Besides, [51] compared three types of speaker models (HMM, circular HMM and supra-segmental HMM) and concluded that the supra-segmental HMM is the most robust against emotion changes. In addition, [51, 52] proposed an adaption approach based on the maximum likelihood linear regression (MLLR) and its feature-space variant, the constrained MLLR (CMLLR). The basic idea is to project the emotional test utterances to neutral utterances by a linear transformation. Then they presented a novel emotion adaptive training (EAT) approach to iteratively estimate the emotion-dependent CMLLR transformations and re-train the speaker models with transformed speech. The results demonstrate that the EAT approach provides significant performance improvements over the baseline system where the neutral enrollment data are used to train the speaker models and the emotional test utterances are verified directly. Finally, some score normalization and transformation approaches [53, 54] have been proposed to improve emotional speaker recognition.

b) Speaking rate

Speaking rate is another high level speaker variable. And it can be useful for discriminating between speakers. Speech events in two utterances - even though they have the same text and are spoken by the same speaker - are seldom synchronized in time. This effect is attributable to the differences in the speaking rates [55]. Researchers [56] have proved that speaking rate has a big impact on speaker verification system performance. Besides, results show that the verification performance for test utterance at normal and slower rates is better than at a faster rate.

The study of speaking rate is originally from speech recognition. In [57], the authors introduced a probabilistic method for estimation the speaking rate in order to choose a recognition model suitable for such speaking rate. From feature domain, [58] presented a speech rate classifier (SRC), which is directly based on the dynamic coefficients of the feature vectors and it is suitable to be used in real time. In addition, [59] proposed three methods to improve the recognition accuracy of fast speech: 1) an implementation of Baum-Welch codebook adaption, 2) the adaptation of HMM state-transition probabilities and 3) the pronunciation dictionaries modified by rule-based techniques.

In speaker recognition, we usually adopt non-linear time alignment or DTW (Dynamic Time Warping) algorithm, which also used in speech recognition, to solve the difference of speaking rates problem. In the text-dependent mode, samples of the same speech events in enrollment and test utterances are compared, usually by establishing a nonlinear time alignment between the utterances. However, it is impossible to conduct time alignment in text-independent case [60]. Therefore, there has not been successful ways to deal with speaking rate issue in text-independent speaker recognition systems.

The difference of speaking rates between enrollment and test utterances can have significant effects on system performance, thus how to apply the relevant methods of speech recognition to settle the speaking rate issue of speaker recognition will be a good research topic in the future.

c) Idiom

Idiom (a person's personal style of word usage) is a high-level inter-speaker characteristic, and can discriminate different speakers who come from different regions or have different educational backgrounds. For instance, we humans can recognize familiar speakers by only a couple of idioms from this speaker. That means through self-learning and training, the human brain can do speaker recognition with the aid of idioms. Scholars in the field of linguistics opened the study on idiom, including idiom principles [61], idiom structure [62], idiomatic expressions [63] and idiom-tagging [64], etc. There have been investigations for automatic speaker recognition with high-level idiom information, including using idiosyncratic word-usage high-level feature [65-67] and idiosyncratic pronunciation feature [68 - 71] etc.

Actually, idiom is not an adverse factor for robust speaker recognition; on the contrary, it can be used as a robust high-level feature to improve robustness. We list it here because if not well dealt with, as a kind of speaking style factor, it may affect the recognition performance.

D. Cross-Lingual

From the results of NIST speaker recognition evaluations in recent years, speaker recognition systems which are mainly developed based on English training data suffer a language gap problem, namely, the performance of non-English trails is much worse than that of English trails. The causes are manifold, such as different characteristics between languages themselves. But the main reason is that a language mismatch between training and testing results in performance degradation in speaker recognition systems. We summarize the previous works in cross-lingual speaker recognition as follows.

a) Training a pooled model from multilingual corpora

Bin *et al*. [72] presented that language mismatch between enrollment and verification data leads to significant performance degradation (between 40% to 49%) . In order to maximize robustness towards language change in test utterances, speaker models were trained with utterances from both languages. Experimental results indicated that this could effectively close performance degradation gap due to language mismatch.

b) Language normalization

In [73], authors proposed two novel algorithms towards spoken language mismatch problem: the first algorithm merges language-dependent system outputs by using Language Identification (LID) scores of each utterance as fusion weights. And in the second algorithm, fusion is done at the segment-level via multilingual Phone Recognition (PR).

c) Feature combination

In [74], authors proposed a combination of features (MFCC and LPCC) methods to obtain a more robust multilingual speaker recognition system. The experimental results indicated that this method could be used for improving the speaker identification performance in multilingual condition with the constraint of limited data.

d) Language factor compensation

Lu *et al*. [75] enrolled in the language factors which are meant to capture the language character of each testing and training speech utterance, and compensation was carried out by removing the language factors in order to shrink the difference between languages. The experimental results showed that the language factor compensation alone can reduce the gap between the performance of English and non-English trails, and the score level combination with eigen-channels can further improve the performance of non-English trails.

E. Aging

It is believed in speaker recognition field that there exists identifiable uniqueness in each voice. At the same time, a question: whether voice changes significantly with time [76]. Similar ideas were expressed in [1, 77], as they argued that a big challenge to uniquely characterize a person's voice was that voice changes over time. Performance degradation has also been observed in the presence of time intervals in practical speaker recognition system [60, 78, 79]. From the pattern recognition point of view, enrollment data (training speaker model) and test utterances for verification are usually separated by some period of time, which poses a mismatch leading to recognition performance degradation.

Therefore, some researchers resorted to several training sessions over a long period of time for dealing with long-term variability issues [80]. In [81], the best recognition performance was obtained when 5 sessions successively separated by at least 1 week were used to define the training set. [82, 83] used a similar method called data augmentation. In this approach, when a positive identification of the candidate speaker is valid, extra data is appended to original enrollment data to provide a more universal enrollment model for this candidate. This approach requires original data to be maintained for re-enrollment. At the same time, speaker-adaptation techniques, such as MAP-adaptation [82, 83] and MLLR-adaptation [84], are proposed to adapt the original model to a new model with new data so that to reduce the

effects of model aging. In [84], experimental results showed that adapting the speaker models on data from the intervening session, Equal Error Rate (EER) on the last two sessions is reduced from 2.5% to 1.7%.

From the score domain, researchers observed that verification scores of true speakers decreased progressively as the time span between enrollment and verification increased, while impostor scores were less affected [85, 86]. Thus a stacked classifier method with an ageing-dependent decision boundary was applied to improve long-term verification accuracy.

From a different direction, Wang [87] aimed to extract speaker-sensitive and time-insensitive information as features. Two methods to determine the discrimination sensitivity of frequency bands were explored: an energy-based F-ratio measure and a performance-driven one. Frequency warping and filter output weighting were performed according to the discrimination sensitivity curves of the whole frequency range. On the time-varying voiceprint database [88], experimental results showed that the proposed features outperformed both MFCCs and LPCCs, and to some extent, alleviated the aging impact on speaker recognition.

IV. Application-oriented Robustness Issues

With the development of speaker recognition technologies, it is used in wide application areas. The main applications of speaker recognition technologies include the following:

A. User Authentication

User authentication is the most obvious application of any biometric authentication techniques. Speaker recognition could be used in commercial transactions as an authentication method combined with some other techniques like face recognition. More recent applications are in user verification for remote electronic and mobile purchases [89]. Alternatively, it could be used for controlling access to computer login, as a "key" to a physical facility, or in border control [90].

B. Public Security and Judicature

Some applications are used in the area of public security and judicature for law enforcement, including parolees monitoring (call parolees at random times to verify they are within the control) and prison call monitoring (validate inmate prior to outbound call) [91]. Speaker recognition technologies have also been used for forensics analysis (proving the identity of a recorder voice to convict a criminal or discharge an innocent in court) [92].

C. Speaker Adaptation in Speech Recognition

Speaker variability is one of the major problems in speech recognition, whereas in speaker recognition it is an advantage. Thus, speaker recognition technology could be used to reduce the speaker variability in speech recognition systems by speaker adaptation (the systems adapt its speech recognizer parameters to suit better for the current speaker, or to select a speaker-dependent speech recognizer from its database) [93].

D. Multi-Speaker Environments

In a multi-speaker environments, searching or tagging speech based on "who spoke when" is one of the more basic components required for dealing with audio recordings, such as recorded meetings or the audio portion of broadcast shows [94]. Three different multi-speaker tasks are recognized: speaker detection (whether a known speaker is present in a multi-speaker recording), speaker tracking (a given speaker's speaking intervals are located in the recording), speaker segmentation (locating the speech intervals of each different speaker) [95].

E. Personalization

The voice-web, voice mail or device customization is becoming more and more popular due to the development in speech technology. Speaker recognition techniques can identify whether the incoming voice is from a known speaker. The system can adapt to a specific speaker's needs and preferences.

The above applications require robust speaker recognition technology. Although, in recent years, the performance of speaker recognition has been considerably high in some restrictive conditions, there are still many issues in practice, including noisy variability, channel variability and intra-individual variation as we mentioned above. All these variations lead to the mismatch between training and test conditions that finally cause severe performance degradations. Therefore, robustness is one of the critical factors that decide the success of speaker recognition in these applications [96].

In addition to the above-mentioned issues, short utterance speaker recognition (SUSR) is also one of the research hotspots in the field of speaker recognition. Most of the speaker recognition methods require a large amount of speech data for training models. SUSR becomes important because of the difficulty in acquiring large amount of appropriate speech. In some situations, only a short utterance such as one or two words is available to recognize the speaker.

Research has shown that the length of test utterance is a great factor that influences the system performance. As early as 1983, researchers from ITT Defense Communication Division have mentioned that for text-independent speaker recognition, both the training and test processes need to have an adequate speech data in order for the accuracy of modeling and recognition. Experimental results showed that the system was able to identify 11 speakers with 96%, 87% and 79% accuracy with test utterance durations of 10, 5 and 3 seconds, respectively [97].

The performances of SUSR were unsatisfactory on these current state-of-the-art speaker recognition systems. [98] illustrated that the impact of restricted test utterance length for GMM-UBM [23] system. These results show that the EER [99] increased sharply from 6.34% to 23.89% when the test utterance is shortened from 20 seconds to 2 seconds. Furthermore, if the length is less than 2 seconds, the EER rose to as high as 35.00%. It has the same phenomena on the basis of JFA [100] and i-vector [29] system, which decrease the number of redundant model parameters to develop more accurate speaker models. [101] examined the effect of shortening the utterances used in subspace training and

speaker model training and scoring. Results suggested that in training the speaker subspace, as much data as possible should be used.

The challenges of SUSR are reflected in the following two aspects [102]:

A. The Mismatch between Training and Test conditions

When the length of test utterance is short, the distribution of speaker information from the short test utterance is unbalanced. Therefore, it only reflects part of the total characteristics space of the speaker. Fig. 1 illustrates the matching degree of different length of test utterance in overall speaker space. Where the outer circle represents the spatial distribution of the target speaker, the shadow in the middle represents the distribution of the test utterance in space. Apparently, when the test data is adequate, it can well reflect the distribution characteristics of the target speaker. Conversely, the test data only reflects a little part of the speaker space and it is hard to make a stable matching validation with the target speaker. The final result is that the test utterance is "not like the target speaker".

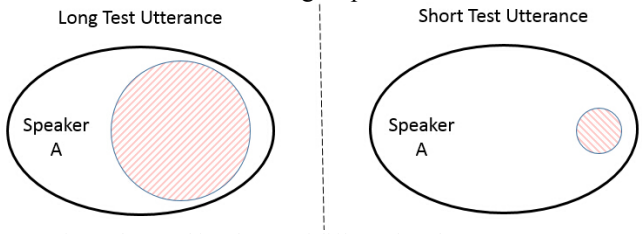

Fig. 1 The matching degree of different length test utterance in speaker space

B. Discriminative Information Inadequate and Confusable

Under the condition of short utterance, the information of test data is inadequate so that it does not provide sufficient distinguishing information and becomes more confusable. As shown in Fig. 2, if the test utterance is long enough, the recognition system can obtain adequate speaker information from the feature vectors, so it has higher discriminative information. On the contrary, if the testing is in the short utterance condition, the speaker information contained in the short utterance is often inadequate and unstable, and the system is not able to exploit the discriminative information. This may result in multi-speaker similarity and reducing system performance. In this situation, the test utterance is "easy to match other speakers".

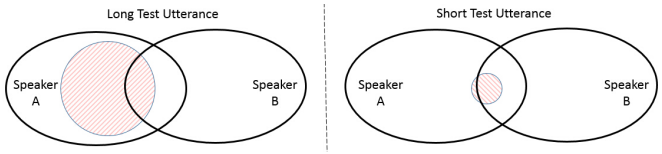

Fig. 2 The confusable degree of different length test utterance in speaker space

In recent years, a multitude of research works have been conducted to address the short utterance issue and made considerate achievement. But overall, the research work on the SUSR task was still in its infancy. Below is an overview of some research directions on SUSR [102].

A. To Select More Discriminative Data

Different parts of a speech signal contain different speaker information, therefore, how to select more effective and discriminative speech data is very important.

[103] described a new method for speaker identification that selectively uses feature vectors for robust decision-making. To reduce the errors due to model overlap, they designed the speaker model in a modified way. Firstly, they classified the feature vectors into two categories: non-overlap and overlap, and then trained two standard speaker models. Next, if the extracted feature vectors were falsely recognized, these features belonged to the overlapped model. In the last, they reconstructed the two models to get non-overlapped and overlapped speaker models. Experimental results showed that this method outperformed GMM and LDA-GMM systems.

Besides, Nosratighods [104] presented a score-based segment selection technique for discarding portions of speech that results in poor discrimination ability in a speaker verification task. Theory is developed to detect the most significant and reliable speech segments based on the probability that test segment comes from a fixed set of cohort models. This approach reduces the effect of acoustic regions of the speech that are not accurately modeled due to sparse training data and makes a decision based only on the segments that provide the best-matched scores. Thus the proposed segment selection technique provides reductions in relative error rate of 22% and 7% in terms of minimum Detection Cost Function (DCF) and EER compared with a baseline used the segment-based normalization, when evaluated on the short utterance of NIST 2002 Speaker Recognition dataset.

In the signal domain, it is not exactly known which information should be extracted from the signal during feature extraction. Actually, each kind of speech feature, such as MFCC [105], PLAR [106] and LPCC [107], represents a sort of partial speaker characteristics. Based on this assumption, a Fisher-voice based feature fusion method incorporating with the Principal Component Analysis (PCA) and the Linear Discriminant Analysis (LDA) was introduced in [108]. There are two steps: first, to de-correlate the concatenated feature vectors (including MFCC, PLAR and LPCC) into individual ones from multiple feature streams; and second, to eliminate the coefficients with redundant and unimportant information by performing PCA and LDA transformation. In conclusion, the main idea of this approach is to select more discriminative feature and provide enough speaker information for SUSR. Compared with the baseline GMM-UBM systems using any single feature and LDA-based fusion method, this proposed method effectively reduces the EER and gives the best performance for text-independent speaker recognition for utterances as short as about 2 seconds.

B. To Train More Accurate Model

As we all know, under the condition of the same training data, different model training methods generate different types of models. The more accurate model can better reflect the personality information of the speaker and increase the

discrimination between different speakers. For SUSR, the process of model training is also an important research subject.

In recent years, Vogt along with his research team from Queensland University published a series of papers and respectively discussed the performance of four state-of-the-art model training methods: GMM-UBM, GMM-SVM, JFA and i-vector [98, 101] [109 - 111]. Vogt expected that in traditional GMM-UBM, most of the parameters in the model do not reflect the speaker information. They motivated the JFA and i-vector to explicitly model and separate the speaker and session contributions. Comparison among GMM-UBM, JFA and i-vector systems on the common subset of the 2008 NIST SRE short2-short3 condition, experimental results showed that when the utterance length was 2 seconds, the value of EERs for GMM-UBM, JFA and i-vector were 23.98%, 22.48% and 21.98%, respectively. In summary, JFA and i-vector have better performance than GMM-UBM for SUSR.

In addition to the above methods, an idea of phoneme specific multi-model method for SUSR was proposed in [112]. They considered that combining the results of speech recognition as *a priori* information to train the speaker model is feasible and efficient. This method consists of a phoneme-class speech recognition and a phoneme-class dependent multi-model speaker recognition stage. During the training procedure, speech recognition is performed to generate a phoneme sequence and all data related to a certain phoneme will be clustered together, then we reconstructed a speaker model depending on these phoneme classes. For recognition, given an utterance, speech recognition will be first performed as in the training procedure to generate a phoneme sequence, each of which will be scored against phoneme specific speaker models of a speaker. After all phonemes have been scored, the score of the utterance against speakers will be obtained. This obviously can change a text-independent task into a text-dependent one. Compared with the baseline GMM-UBM system, this proposed method can achieve a relative EER reduction of 38.60% for text-independent SUSR with the utterance length less than 2 seconds.

C. Better Algorithms for Scoring

The classical decision method of GMM-UBM is based on the log-likelihood ratio computation. That is to calculate the difference between UBM log-likelihood and speaker's log-likelihood as an evaluation criterion. Malegaonka considered that this is a unilateral scoring (ULS) [113, 114]. That is, when the models built using speech from a speaker (speaker A) are matched against speech from another speaker (speaker B), they may not return high likelihoods whilst speech from speaker A matched against the models built using speech from speaker B giving high likelihoods. Thus it does not reflect the similarity degree between speakers, especially for SUSR. So a weighted bilateral scoring (WBLS) method has been proposed and investigated. In practice, the enrollment utterances of speakers are usually longer in duration than the test utterance. As a result, the speaker models built using the enrollment speech are more reliable than those built using the test tokens. Therefore, they proposed an approach to deal with this imbalance in the quality of speaker models by appropriately weighting the forward and reverse similarity scores before combining them. At last, it is demonstrated experimentally that the EER can be relatively reduced about 20% by using the proposed scoring method comparing with the ULS.

V. CONCLUSIONS

This paper presents an overview of speaker recognition technologies with an emphasis on dealing with robustness issues. We summarize the speaker recognition robustness issues from three categories, including environment-related robustness issues, speaker-related robustness issues and application-oriented robustness issues. We give overview of existing technologies to deal with robustness issues in each category and present potential research focuses in the future.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant No.61271389/61371136 and the National Basic Research Program (973 Program) of China under Grant No.2013CB329302.

REFERENCES

[1] S. Furui, "Recent Advances in Speaker Recognition," *Pattern Recognition Letters*, 18, (1997), 859-872.

[2] J. P. Campbell, "Speaker Recognition: A Tutorial," *Proceedings of the IEEE*, 85, 9, (1997), 1437-1462.

[3] S. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. Audio, Speech and Language Processing 14*, 5 (September 2006), 1557–1565.

[4] T. Kinnunen and H. Li, "An Overview of Text-Independent Speaker Recognition: From Features to Supervectors," *Speech Communication 2010*, 52 (2010), 12-42.

[5] A. Rosenberg and F. Soong, "Recent research in automatic speaker recognition," in *Advances in Speech Signal Processing*, S. Furui and M. Sondhi, Eds. New York, NY: Marcel Dekker, 1992, ch. 22, pp. 701–738.

[6] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust. Speech Signal Processing*, 1981. 29(2):254-272.

[7] D. A. Reynolds, "Channel robust speaker verification via feature mapping," *In ICASSP*, 2003, (2): 53-56.

[8] G. R. Doddington, M. A. Przybocki, A. F Martin, D. A Reynolds, "The NIST speaker recognition evaluation–Overview, methodology, systems, results, perspective," *Speech Communication*, 2000, 31(2): 225-254.

[9] The NIST Year 2008 Speaker Recognition Evaluation Plan, *http://www.itl.nist.gov/iad/mig//tests/spk/2008/sre08_evalplan_release4.pdf.*

[10] The NIST Year 2010 Speaker Recognition Evaluation Plan, *http://www.itl.nist.gov/iad/mig//tests/spk/2010/NIST_SRE10_evalplan.r6.pdf.*

[11] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1979, 27:113-120.

[12] M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *ICASSP*, 1979, 208-211.

[13] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, 1994. 2(4): 578-589

[14] A. Kocsor, L. Toth, A. Kuba, K. Kovacs, M. Jelasity, T. Gyimothy, J. Csirik, "A comparative study of several feature transformation and learning methods for phoneme classification," *International Journal of Speech Technology*, 2000. 3(3): 263-276.

[15] R. G. Lomax and D. L. Hahs-Vaughn, "Statistical concepts: a second course," *Lawrence Erlbaum Associates*, 2007.

[16] G. Saon, M. Padmanabhan, R. Gopinath, S. Chen, "Maximum likelihood discriminant feature spaces," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 2000. 2: 1129-1132.

[17] Gales, J. F. Mark and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Transactions on Speech and Audio Processing*, 1996, 4(5): 352-359.

[18] P. Renevey and A. Drygajlo, "Statistical estimation of unreliable features for robust speech recognition," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP 2000)*, Istambul, Turkey, 2000, 1731-1734.

[19] D. Zhu, B. Ma, H. Li, and Q. Huo, "A generalized feature transformation approach for channel robust speaker verification," in *Proc. ICASSP'07*, vol. 4, 2007, pp. 61–64.

[20] C. Vair, D. Colibro, and P. Laface, "Channel factors compensation in model and feature domain for speaker recognition," in *Proc. Odyssey Speaker and Language Recognition Workshop*, 2006, pp. 1–6.

[21] L. P. Heck and M. Weintraub, "Handset-dependent background models for robust text-independent speaker recognition," in *Proc. ICASSP'97*, vol. 2, 1997, pp. 1071–1074.

[22] R. Teunen, B. Shahshahani, and L. Heck, "A model-based transformational approach to robust speaker recognition," in *Proc. ICSLP'00*, 2000, pp. 495–498.

[23] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[24] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 42–54, 2000.

[25] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," In *Proc. Speaker Odyssey 2001 conference*, June 2001, pp.213-218

[26] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[27] A. Solomonoff, C. Quillen, and W. M. Campbell, "Channel compensation for SVM speaker recognition," in *Proc. Odyssey Speaker and Language Recognition Workshop*, 2004, pp. 57–62.

[28] D. Matrouf, N. Scheffer, B. Fauve, and J.-F. Bonastre, "A straightforward and efficient implementation of the factor analysis model for speaker verification," in *Proc. Interspeech' 07*, 2007, pp. 1242–1245.

[29] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[30] A. O. Hatch, S. S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *INTERSPEECH' 06*, 2006.

[31] S. Ioffe, "Probabilistic linear discriminant analysis," in *ECCV 2006*, 2006, pp. 531–542.

[32] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *ICCV'07. IEEE*, 2007, pp. 1–8.

[33] M. McLaren and D. A. van Leeuwen, "Source-normalised-and-weighted LDA for robust speaker recognition using i-vectors," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 5456–5459, 2011.

[34] S. P. Kishore and B. Yegnanarayana, "Speaker verification: minimizing the channel effects using autoassociative neural network models," in *Proc. ICASSP'00*, vol. 2, 2000, pp. 1101–1104.

[35] C. Huang, T. Chen, S. Li, E. Chang and J.-L. Zhou, "Analysis of Speaker Variability," *Microsoft Research*, China, 2001.

[36] S. Cumani, O. Glembek, N. Brummer, E. de Villiers, P. Laface, "Gender independent discriminative speaker recognition in i-vector space," *ICASSP,* 2012.

[37] M. McLaren and D. A. van Leeuwen, "Gender-independent speaker recognition using source normalization," *in Proc. ICASSP*, 2012, pp.4373-4376.

[38] R. G. Tull and J. C. Rutledge, "Analysis of 'cold-affected' speech for inclusion in speaker recognition systems," *Acoustical Society of America, The Journal of the Acoustical Society of America, Volume 99*, 2549, 1996.

[39] R. G. Tull and J. C. Rutledge, "'Cold Speech' for Automatic Speaker Recognition," *Acoustical Society of America 131st Meeting Lay Language Papers*, May, 1996.

[40] R. G. Tull, J. C. Rutledge and C. R. Larson, "Cepstral analysis of "cold-speech" for speaker recognition: A second look," *Acoustical Society of America, The Journal of the Acoustical Society of America*, Volume 100, 2760, 1996.

[41] R. G. Tull, "Acoustic analysis of cold-speech: implications for speaker recognition technology and the common cold," *Phd thesis*, Northwestern University, 1999.

[42] O. W. Kwon, K. Chan, J. Hao, T. W. Lee, "Emotion recognition by speech signals," *INTERSPEECH*, 2003.

[43] B.-H. Juang, "Speech recognition in adverse environments," *Computer speech & language*, 1991, 5(3): 275-294.

[44] R. Lippmann, E. Martin, D. B. Paul, "Multi-style training for robust isolated-word speech recognition," *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'87. IEEE*, 1987, 12: 705-708.

[45] F.-H. Bie, D. Wang, T. F. Zheng, R. Chen, "Emotional speaker verification with linear adaptation," *Signal and Information Processing (ChinaSIP), 2013 IEEE China Summit & International Conference on,. IEEE*, 2013: 91-94.

[46] E. Zetterholm, "Prosody and voice quality in the expression of emotions," in *Proc. ICSLP'98*, 1998, pp. 109–113.

[47] T. Wu, Y.-C. Yang, and Z.-H. Wu, "Improving speaker recognition by training on emotion-added models," in *Proc. Affective Computing and Intelligent Interaction*, 2005, pp. 382–389.

[48] C. Pereira and C. Watson, "Some acoustic characteristics of emotion," in *Proc. ICSLP'98*, 1998, pp. 927–930.

[49] K. R. Scherer, T. Johnstone, G. Klasmeyer, and T. Banziger, "Can automatic speaker verification be improved by training the algorithms on emotional speech?" in *Proc. ICSLP'00*, 2000, pp. 807–810.

[50] K. R. Scherer, D. Grandjean, T. Johnstone, G. Klasmeyer, and T. Banziger, "Acoustic correlates of task load and stress," in *Proc. ICSLP'02*, 2002, pp. 2017–2020.

[51] I. Shahin, "Speaker identification in emotional environments," *Iranian Journal of Electrical and Computer Engineering,* vol. 8, no.1, pp. 41–46, 2009.

[52] F.-H. Bie, D. Wang, T. F. Zheng, J. Tejedor, R. Chen, "Emotional adaptive training for speaker verification," *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific. IEEE*, 2013: 1-4.

[53] W. Wu, T. F. Zheng, M.-X. Xu, and H.-J. Bao, "Study on speaker verification on emotional speech," in *Proc. Interspeech'06*, 2006, pp.2102–2105.

[54] Z.-Y. Shan and Y.-C. Yang, "Learning polynomial function based neutral-emotion GMM transformation for emotional speaker recognition," in *Proc. ICPR'08*, 2008, pp. 1–4.

[55] B. S. Atal, "Automatic recognition of speakers from their voices," *Proceedings of the IEEE*, 1976, 64(4): 460-475.

[56] T. Matsui, S. Furui, "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/ continuous HMM's," *Speech and Audio Processing, IEEE Transactions on*, 1994, 2(3): 456-459.

[57] H. Yasuda and M. Kudo, "Speech rate change detection in martingale framework," in *Proc. ISDA*, 2012, pp.859-864.

[58] F. Martinez, D. Tapias, J. Alvarez, "Towards speech rate independence in large vocabulary continuous speech recognition," *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on. IEEE*, 1998, 2: 725-728.

[59] M. A. Siegler, R. M. Stern, "On the effects of speech rate in large vocabulary speech recognition systems," *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on. IEEE*, 1: 612-615.

[60] F. Soong, A. E. Rosenberg, L. R. Rabiner, B. H. Juang "A vector quantization approach to speaker recognition," in *Proc. of ICASSP 1985*, vol. 10, pp. 387-390, Florida, 1985.

[61] B. Erman, B. Warren, "The idiom principle and the open choice principle," *TEXT-THE HAGUE THEN AMSTERDAM THEN BERLIN*, 2000, 20(1): 29-62.

[62] A. Makkai, "Idiom structure in English" *Walter de Gruyter*, 1972.

[63] C. Cacciari, S. Glucksberg, "Understanding idiomatic expressions: The contribution of word meanings," *Advances in Psychology*, 1991, 77: 217-240.

[64] G. Leech, R. Garside, M. Bryant, "CLAWS4 the tagging of the British National Corpus," *Proceedings of the 15th conference on Computational linguistics-Volume 1. Association for Computational Linguistics*, 1994: 622-628.

[65] G. Doddington, "Speaker Recognition based on Idiolectal Differences between Speakers,"*Eurospeech, Vol. 4, pp. 2517-2520, 2001.*

[66] S. Kajarekar, L. Ferrer, E. Shriberg, K. Sonmez, A. Stolcke, A. Venkataraman, and J. Zheng, "SRI's 2004, NIST Speaker Recognition Evaluation System," In *Proc. ICASSP. 2005*.

[67] G. Tur, E. Shriberg, A. Stolcke, S. Kajarekar, "Duration and Pronunciation Conditioned Lexical Modeling for Speaker Verification," In *Proc. of Interspeech, 2007*.

[68] W. Andrews, M. Kohler, J. Campbell, J. Godfrey, and J. Hernandez-Cordero, "Gender-dependent Phonetic Refraction For Speaker Recognition," *ICASSP 2002*.

[69] J. Navratil, Q. Jin, W. Andrews, and J. Campbell, "Phonetic Speaker Recognition Using Maximum Likelihood Binary Decision Tree Models," *ICASSP 2003*.

[70] Q. Jin, J. Navratil, D. Reynolds, J. Campbell, W. Andrews, J. Abramson, "Combining Cross-Stream And Time Dimensions In Phonetic Speaker Recognition," *ICASSP 2003*.

[71] A. Hatch, B. Peskin, A. Stolcke, "Improved Phonetic Speaker Recognition Using Lattice Decoding," In *Proc. ICASSP. 2005*.

[72] M. Bin, and H.-L. Meng, "English-Chinese bilingual text-independent speaker verification," *Acoustics, Speech, and Signal Processing, 2004. Proceedings (ICASSP'04). IEEE International Conference on.* Vol. 5. IEEE, 2004.

[73] M. Akbacak, J. H. Hansen, "Language normalization for bilingual speaker recognition systems," *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. IEEE, 4: IV-257-IV-260.

[74] B. G. Nagaraja, H. S. Jayanna, "Combination of Features for Multilingual Speaker Identification with the Constraint of Limited Data," *International Journal of Computer Applications*, 2013, Vol.70 (6), pp.1-6.

[75] L. Lu, Y. Dong, X. Zhao, J. Liu, H. Wang, "The effect of language factors for robust speaker recognition," *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009: 4217-4220.

[76] L.G. Kersta, "Voiceprint Recognition," *Nature*, No. 4861, pp. 1253-1257, December 1962.

[77] J. Bonastre, F. Bimbot, L. Boe, J. P. Campbell, "Person authentication by voice: a need for caution," *Proc. of Eurospeech 2003*, pp. 33-36, Geneva, 2003.

[78] T. Kato and T. Shimizu, "Improved speaker verification over the cellular phone network using phoneme-balanced and digit-sequence preserving connected digit patterns," in *Proc. of ICASSP 2003*, Hong Kong, 2003.

[79] M. Hebert, "Text-dependent speaker recognition," *Springer Handbook of Speech Processing*, Springer-Verlag: Berlin, 2008.

[80] F. Bimbot, J. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, iss. 4, pp. 430-451, 2004.

[81] J. Markel and S. Davis, "Text-independent speaker recognition from a large linguistically unconstrained time-spaced database," *IEEE Transactions on Acoustics*, *Speech*, *and Signal Processing*, Volume ASSP-27, No. 1, pp. 74-82, February 1979.

[82] H. Beigi, "Effects of time lapse on speaker recognition results," *Proc. of 16th International Conference on Digital Signal Processing*, pp. 1-6, 2009.

[83] H. Beigi, "Fundamentals of speaker recognition," *New York Springer*, 2010.

[84] L. Lamel and J. Gauvin, "Speaker verification over the telephone," *Speech Communication*, Volume 2000, Issue 31, pp. 141-154, 2000.

[85] F. Kelly and N. Harte, "Effects of long-term ageing on speaker verification," *Biometrics and ID Management*, Volume 6583 of *Lecture Notes in Computer Science*, pp. 113-124, Springer Berlin/Heidelberg, 2011.

[86] F. Kelly, A. Drygajlo, and N. Harte, "Speaker verification with long-term ageing data," in *Proc. of 5th IAPR International Conference on Biometrics*, New Delhi, 2012.

[87] L.-L. Wang, X.-J. Wu, T. F. Zheng and C.-H. Zhang, "An Investigation into Better Frequency Warping for Time-Varying Speaker Recognition," *APSIPA ASC*, 2012.

[88] L.-L Wang and T. F. Zheng, "Creation of time-varying voiceprint database," *Proc. of O-COCOSDA 2010*, Kathmandu, 2010.

[89] D. A. Reynolds, "Automatic Speaker Recognition Using Gaussian Mixture Speaker Models," *The Lincoln Laboratory Journal*, 1995.

[90] T. Kinnunen, "Spectral Features for Automatic Text-Independent Speaker Recognition," *LICENTIATE'S THESIS*, Dec. 2003.

[91] D. A. Reynolds, L. Heck, "Speaker Verification: From Research to Reality," *ICASSP,* 2001.

[92] P. Rose, "Forensic Speaker Identification," *Taylor & Francis*, London, 2002.

[93] R. Kuhn, J.C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Transactions on Speech and Audio Processing*, 8(6):695–707, 2000.

[94] R. B. Dunn, D. A. Reynolds, and T. F. Quatieri, "Approaches to Speaker Detection and Tracking in Conversational Speech," *Digital Signal Processing*, 2000, pp.93-112.

[95] A. Martin, M. Przybocki, Speaker recognition in a multi-speaker environment, In *Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001)* (Aalborg, Denmark, 2001), pp. 787 790.

[96] Q. Jin, "Robust Speaker Recognition," *Carnegie Mellon University Pittsburgh*, Jan. 2007.

[97] K. Li and Jr. E. Wrench, "An approach to text-independent speaker recognition with short utterances," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8: 555-558, 1983.

[98] R. Vogt, S. Sridharan and M. Mason, "Making confident speaker verification decisions with minimal speech," *IEEE Trans. on ASLP*, vol. 18, no. 6, pp. 1182-1192, 2010.

[99] NIST Speaker Recognition Evaluation Plan, Online Available *http://www.nist.gov/speech/tests/sre/*.

[100] P. Kenny, P. Dumouchel, "Experiments in Speaker Verification using Factor Analysis Likelihood Ratios," in *Proceedings of Odyssey04 - Speaker and Language Recognition Workshop*, Toledo, Spain, 2004.

[101] R. Vogt, B. Baker, and S. Sridharan, "Factor analysis subspace estimation for speaker verification with short utterances," in *Interspeech*, Brisbane, 2008.

[102] C.-H. Zhang, "Research on Short Utterance Speaker Recognition," *Phd thesis, Tsinghua University*, April 2014.

[103] S. Kwon and S. Narayanan, "Robust speaker identification based on selective use of feature vectors," *Pattern Recognition Letters*, 28 (1): 85-89, 2007.

[104] M. Nosratighods, E. Ambikairajah, J. Epps and M. J. Carey, "A segment selection technique for speaker verification," *Speech Communication*, 52 (9): 753-761, 2010.

[105] S. B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on ASLP*, vol. 28, no. 4, pp. 357–366, 1980.

[106] D. Chow and W. H. Abdulla, "Robust speaker identification based perceptual log area ratio and Gaussian mixture models," *Interspeech*, 2004.

[107] P. Premakanthan and W. B. Mikhael, "Speaker verification recognition and the importance of selective feature extraction: review," *MWSCAS*, vol. 1, 57-61, 2001.

[108] C.-H. Zhang and T. F. Zheng, "A fishervoice based feature fusion method for short utterance speaker recognition," *IEEE China Summit and International Conference on Signal and Information Processing, ChinaSIP*, 2013.

[109] M. McLaren, R. Vogt, B. Baker and S. Sridharan, "Experiments in SVM-based speaker verification using short utterances," *A Speaker Odyssey-The Speaker Recognition Workshop*: 83-90, 2010.

[110] M. McLaren, R. Vogt, B. Baker and S. Sridharan, "Data-driven background dataset selection for SVM-based speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, 18 (6): 1496-1506, 2010.

[111] A. Kanagasundaram, R. Vogt, D. B. Dean, S. Sridharan and M. W. Mason, "i-vector based speaker recognition on short utterances," *International Speech Communication Association (Interspeech)*: 2341-2344, 2011.

[112] C.-H. Zhang, X.-J. Wu, T. F. Zheng and L.-L. Wang, "A K-phoneme-class based multi-model method for short utterance speaker recognition," *The 4th Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference, APSIPA ASC*, 2012.

[113] E. S. Parris and M. J. Carey, "Multilateral techniques for speaker recognition," *International Conference on Spoken Language Processing (ICSLP)*, 1998.

[114] A. Malegaonkar, A. Ariyaeeinia, P. Sivakumaran and J. Fortuna, "On the enhancement of speaker identification accuracy using weighted bilateral scoring," *IEEE International Carnahan Conference on Security Technology (ICCST)*: 254-258, 2008.