Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 14 (2020) Pp-419-426

BBRC
Bioscience Biotechnology Research Communications

Brief Review of Short Utterance Speaker Verification Systems

Asmita Nirmal[1] and Deepak Jayaswal[2]

[1]*Assistant Professor, Department of Electronics Engineering, Datta Meghe College of Engineering, Navi Mumbai, India*

[2]*Professor, Department of Electronics and Telecommunication Engineering, St. Francis Institute of Technology, Mumbai, India*

ABSTRACT

Due to technological improvements many methods have been proposed for speaker verification. While performance is satisfactory with large amounts of speech data, there is significant degradation in performance with short utterances. Many research works have been carried out to handle short utterance issue of the speaker verification systems used in real-world scenario. In this paper we primarily emphasis on the survey of different feature extraction methods for text-independent speaker verification. We first briefly review conventional systems to show its progress. In this work, we present a brief review of features that are used to capture speaker information at different analysis lengths of speech utterance. We also put the major findings and challenges of this research feview in a nutshell.

KEY WORDS: SPEAKER VERIFICATION, SHORT UTTERANCES, FEATURE EXTRACTION.

INTRODUCTION

Speech signal is a main source of speaker specific information. Jain et al., 2004 have shown that besides containing the information related to behavioral aspects speech also contains information of speaker's speech production system. This speaker specific information conveyed by speech signal motives us to use speech signal as a biometric trait. Speaker recognition is the process of automatically recognizing a speaker from his/her speech utterances. Speaker recognition has two categories of tasks: verification and identification. From the speaker recognition systems reviewed by Cambell,1997 speaker identification (SI) is the process of comparing the input speech signal with the models of registered speakers. In contrast, speaker verification (SV) is a process of verifying a claimed identity of a person from his/her speech. Depending on the text contents of the speech speaker recognition systems are categorized into text-dependent (TD) and text-independent (TI) systems. These approaches are used in the studies proposed by Rodriguez –Linares et al., 1998 and Mengistu et al., 2017 respectively. The text-dependent systems have same text content for training and testing phase. Unlike TD systems, TI systems have no control over the speech contents. In TI systems text content for training and testing phase can be different.

a. Motivation: During the last few decades, the use of speech and speaker recognition techniques is increased in smartphones, and various hhandheld devices. Almost all of these devices are used in applications subject to noisy conditions. Along with this the channel variations introduced from the handset and/or microphone devices are also of major concern. Many solutions have been proposed to provide robustness in such practical situation. The performance of existing SV systems have been found satisfying when sufficient amount of speech data is available as shown in the approach proposed by

ARTICLE INFORMATION

**Corresponding Author:* nirmalasmita2607@gmail.com

Received 18th Oct 2020 Accepted after revision 24th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

Clarivate Analytics

crossref

NAAS Journal Score 2020 (4.31)

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.14/96

Kounoudes et al.,2006. However in most of the realistic cases such as forensic applications proposed by Jayanna et al., 2009.,it is hardly impracticable to get sufficient data and that also covering intra-speaker variability to mitigate the effects of the realistic environment. In case of access control systems deployed in banking applications, the average input test speech is limited just for few seconds.

In both of these scenarios, very less amount of feature vectors will be available for enrollment and evaluation, which causes poor speaker modeling and give unpredictable decision of verification. Hence, it is important to consider the effects of the real-life environment.Further to have reliable performance in practical applications one should take into consideration the problem of limited speech data availability. Concerned to these issues different approaches for speaker verification system have been developed for short utterance based speaker verification by Fatima et al.,2012, Matza et al.,2011. In this paper, we make a broad survey of short utterances SV systems considering the studies from recent research. From this review we also give summary of the major findings, issues and various solutions in short utterance point of view.

b. Organization: In this paper the key emphasis is on review of short utterance TI-SV systems. This paper is organised as follows: First the conventional speaker verification systems are reviewed to show development in this area . Next to this the detailed literature review of features extraction methods which are suitable for the short utterances based SV systems is done.Then, major findings from the review followed by different future opportunities and challenges to be handled in this area are discussed. Lastly, the conclusions for the reviewed work are drawn.

2. Basic Components of Speaker Verification System: Before reviewing various text-independent SV research works in short utterance framework, we first detail basic components of SV system. The feature extraction module extracts feature vectors from the raw input speech to form voiceprint of a speaker. During the enrollment phase, a speaker model is trained using extracted features obtained from the feature extraction module Then the trained model is stored into the database. During the verification phase, the features extracted from test speech are compared against the claimed model to compute a similarity score. Finally, this similarity score is used by a decision module to accept or reject the authencity of input test speech. From frame duration point of view, SV systems divide features extraction methods into three categories: source features, short term features and high level features.

Source features are computed from short duration frames of 3-5 milliseconds. An approach proposed by S.R.M.

Prasanna et al., 2006 use sourch features for representing glottal flow information of a speaker .Short term feature analysis use speech frames of about 20–30 milliseconds. These features convey vocal tract information of the speaker. Existing verification systems commonly use vocal tract features because of their less computational complexity. High level features used by Doddington ,2001 capture conversation level information of speakers. These features use frames of 100-300 milliseconds in duration and can capture information such as speaker's word usage, speaking style. High level features are comparatively robust but can be spoofed easily. Selection of particular type of features depends on type of application, computational complexity and amount of available speech data as per the study made by Reynolds,2003.

From the last five decades SV systems are advanced significantly from models based on direct speaker specific features to Gaussian mixture models (GMM) based models proposed by Reynolds et al., 1995.The main reason of progress in the speaker recognition area is the development of various session compensation methods for both GMM and support vector machines (SVM).The detail study of this can be found in the study made by Campbell,2006.With the aim of adapting the GMM-based acoustic model to new operating conditions to compensate intersession variability, a speaker independent model known as universal background model was proposed by Reynolds et al., et al.,2000 . Then study made by Kenny et al.,2007 has extended GMM-UBM model to latent variable based method known as Joint factor analysis (JFA).

In this study supervector space were developed to solve the session variability issue. Combination of JFA compensation and Gaussian supervector SVMs was studied by Dehak et al., 2008. Recent research proposed by Dehak et al., 2011 has introduced the i-vector based speaker recognition. The idea of i-vector is initiated from a JFA that models speaker and channel subspaces separately. In contrast to JFA, i-vector use a single sub-space to represents both speaker and channel variability. Over the past few years the i-vector based on probabilistic linear discriminant analysis (PLDA) modeling which was proposed by Prince et al., 2007 have been developed in SV field[19]. Now a days deep neural network (DNN) are widely used to extract speaker specific information. Lei, et al., 2014 have suggested the use of DNN in a i-vector framework to capture pronunciation patterns of a speaker . The results obtained from this approach have shown performance improvements over the conventional GMM- UBM framework.

3. Review of different feature extraction methods for short utterance speaker verification systems: From the recent few decades many methods are developed to handle the short utterance issue at different levels of speaker verification system. In this section we primarily explore various research studies that are carried out at feature level. List of of different methodologies used in various research studies is shown in Table 1.

a. Low level Features: The most commonly used vocal tract feature in conventional SV systems is Mel-frequency cepstral coefficients (MFCC).However, It is shown in the approach proposed by Kanagasundaram ,2011 performance drops significantly when these features are used for short utterances SV. In this consent features that are less sensitive to lexical content of speech should be focused more. In addition to this the use of complementary information captured by different low level features should be used. Fusion of systems using different set of low level features proposed by Hosseinzadeh et al., 2007 helps in improving the performance. The local covariance features are based on eigen-structure of covariance matrix. Unlike cepstral and delta features, covariance matrix captures uncertainty information. The authors , Sahidullah et al., 2016 have investigated the use of individual as well as fusion of features such as frequency domain linear prediction (FDLP) , mean Hilbert envelope coefficient (MHEC) and power-normalized cepstral coefficients(PNCC) . Further, a new feature set known as constant Q cepstral

coefficient (CQCC) derived from constant Q transform (CQT) is recommended in . CQCC features characterize: the human auditory system. The detail study of FDLP,MHEC, PNCC,CQCC can be found in the approaches proposed by Athineos et al., 2007, Sadjad et al., 2015,Kim et al., 2012 Todisco et al.,2016 respectively. The goal of CQT is similar to RASTA filtering. It focuses on extracting the information pertinent to the articulation rate of the speaker. In contrast to RASTA, however, the CQCC filter bank is adaptive to speech utterance. Further the complementary information captured by different features is explored through the use of Robust Speaker Recognition (RSR 2015) dataset.

b.Source features: The subsequent work proposed by Patil et al.,2013 suggests the use of Liljencrants–Fant (LF) parameters to characterize the glottal flow derivative (GFD) by locating the glottal closing and opening instants. Explicit and implicit modeling of glottal.

information along with their comparison is done in this approach. Explicit approach is more suitable for verification task as it captures small intra-speaker variation. In contrast, implicit approach is found to be useful for SI as it captures large inter-speaker variation. Individually both implicit and explicit methodology signifies the speaker characteristics complimentary to the conventional vocal tract based approach. Following this Chen et al.,2013 have proposed a noise separation method motivated constrained non-negative matrix factorization (CNMF) of short utterances . This method uses difference detection and discrimination (DDADA) algorithms to categorize speech into high quality and low quality speech.

The features from different quality speech are then used in the conventional GMM-UBM framework. Li et al.,2015 have suggested the use of the multi-resolution time frequency feature (MRTF). This study is based on the idea that speaker specific information might be available in the spectrogram calculated at different time frequency scales. Two dimensional cosine transform of spectrogram matrix calculated at different scales is used for forming the feature set. Systems with this feature set have shown superior performance when tested on National Institute of Standards and Technology Speaker Recognition Evaluation (NIST SRE) 2008 corpus. Fusion of amplitude and phase-information is proposed byAlam et al.,2015. For the amplitude- related different cepstral features are considered whereas for phase related features, modified group delay and all-pole group delay, linear prediction residual are considered. The average fused system has shown EER improvement.

A new feature set based on instantaneous frequency cosine coefficient (IFCC) free from phase warping issue is suggested by Vijayan et al., 2016. The improved results obtained from the fusion of IFCC feature with the MFCC and FDLP features shows complementary nature of information captured by individual features. Recent research suggests the use of deep neural network (DNN) based speaker verification. In one of approach proposed by Guo et al.,2016. DNN is used as a regression model which transforms filter-bank coefficients of the speech signal to the associated sub-glottal features. A distinct approach, mainly useful for degraded condition is proposed by Bharathi et al.,2013.It measures amount of non-stationarity of speech signal using amplitude and frequency modulation concept. For doing this study it has used Texas instruments and Massachusetts institute of technology (TIMIT) database. Further, more advanced analysis technique motivated by work proposed by proposed by Ambikairajah et al.,2007 uses empirical mode decomposition (EMD) for feature extraction. It has demonstrated that this approach captures the information complementary to that of vocal tract and source excitation features.

c.High level features: Mary et al.,2008 have [36] proposed use of prosodic features for speaker verification. It is based on the assumption that prosody is related to linguistic units such as syllables. The syllables are extracted by detecting the vowel onset points which is primarily helpful when explicit syllable boundaries are not easily obtainable. Prosodic features formation using pitch and energy contours of speech is studied by Dehak et al.,2007. Significant improvement is achieved when these prosodic features are combined with MFCC features for GMM based modeling over the conventional JFA based modeling approach. One can distinguish speakers just by listening the one who is familiar than the one who is not. Based on this fact the work studied by proposed by Doddington et al.,2001 use various idiolectal dissimilarities of speakers. The results achieved are very encouraging but feasible with a sufficient amount of training data. Another innovative approach proposed by Andrew et al., 2002 has used phone sequences from multiple languages to create gender-dependent phone models. It is observed that this strategy helps in reducing cross-talk from input speech.

Table 1. Review of short utterance speaker verification research

Year of Publication	Methodology	Database
2001	Bigram statistics from familier speaker characteristics such as speaker specific words	SwitchBoard
2002	Extract phone sequences using phonetic recognizers for speaker modeling	Switchboard
2006	Conditional pronunciation modeling of articulatory features	SPIDRE ,Switchboard
2007	Vocal tract and excitation feature extraction from LPC based group delay Prosodic feature extraction by fitting pitch and energy contours with Legendre polynomial expansions	NIST SRE 2001 NIST SRE 2006
2008	Prosodic features associated to linguistic units such as syllables directly extracted from speech	NIST SRE 2003
2009	Multiple frame size and frame rate instead of single frame size and rate	TIMIT
2011	Match the pitch and MFCC contours using dynamic time warping	NIST SRE 2008
2012	Auto-encoder bottleneck feature ASCCD	
2013	Glottal closing and opening instants are located using Liljencrants–Fant parameters	NIST SRE 1999 , 2003
2014	Multiresolution analysis of speech spectrogram Phone discriminant and speaker discriminant DNN as deep features	NIST SRE 2008 RSR2015
2015	Phonetic contents Amplitude- related cepstral features and phase related modified group delay and all-pole group	RSR2015 and NIST SRE 2010 NIST SRE 2008 and 2010.
2016	delay, linear prediction residual are fused togather Tackle the phase warping issue using Instantaneous frequency cosine coefficients	NIST SRE 2010
	Estimation of subglottal acoustic features based on DNN Eigen-structure analysis of covariance matrix of local short term features Constant Q cepstral coefficients (CQCC) features inspired by the human auditory system	NIST SRE 2008 NIST SRE 2001, 2008 ,2010 and RSR2015 NIST SRE RSR2015
2017	Voice quality features motivated by a psycho-acoustic model	NIST SRE 2010
2018	Patterson-Holdsworth Meddis hair cell model Human based subjective evaluations and machine based evaluations using the high level speaker characteristics	TIMIT UCLA
2019	Weighted sum of phoneme variations	NIST SRE 2010

The work studied by Leung et al.,2006 makes use of speaker pronunciation of the speakers who are from different educational background, and different accent. Further this study has also investigated the relationship between articulatory features and phoneme patterns of speakers using conditional pronunciation model. An analysis of deep features in a Tandem method for speaker verification is studied by FU et al.,2014. Phone discriminant and speaker discriminant DNN are combined with conventional acoustic features and applied in a GMM-UBM framework. Another strategy proposed by Sainath et al.,2012 is based on neural network (NN) bottleneck features is experimented on Annotated Speech Corpus of Chinese Discourse (ASCCD). Here, a constant number of hidden units are used to predict output targets. Then, auto-encoder bottleneck feature set is formed using dimensionality reduced output target probabilities. Another system which makes use of bottleneck features for studying the language independent speaker verification can be explored in the

study made by Fatima et al., 2012. Another study made by have used different vowel categories as high level features for SV in a GMM-UBM framework.

The work made by Scheffer et al., 2012 is a study to match the content from a speaker's enrollment data with the test data content. Matching of the contents is done at the statistical level. Inspired by a psycho-acoustic model proposed by Park et al., 2017 make use of voice quality features. The study proposed by Park et al., 2018 compares SV performance human based subjective evaluations and machine based evaluations using the high level speaker characteristics like speaking style. It is shown that evaluations done by humans and systems based on University of California Los Angeles (UCLA) speaker variability database are dissimilar.The system performance can be enhanced by investigating more advanced relation between different acoustic features and perceptual features. Paulose et al., 2017 suggests the use of inner hair cell (IHC) coefficients based on the physiological variations of the mammalian outer hearing system.

A significant rise in the performance is achieved when the IHC features are combined with pitch and formants. The phonetic influence of short utterances is analyzed in the study made by Vinals et al., 2019 .It is shown that weighted sum of phoneme influences is representative of the speaker specific auditory system component to some extent. When the weights fluctuate from the required weight distribution, they do not contribute to the speaker specific information and hence cause performance degradation. A system detecting whether the input is a authentic or a recorded speech is used in the work done by Villalba et al., 2011 to avoid false acceptance. In this system short testing utterances are formed by cutting and pasting the speech segments of registered speaker's utterances. The spoofing attack is detected by matching the pitch and MFCC contours of the enrollment and test segments using dynamic time warping. This study is to avoid spoofing of speaker verification system essentially in the presence of replay attack.

4. Major findings and Future Opportunities and challenges: In this paper we have mainly briefed the research solutions and challenges considering the feature extraction approaches to be eployable in the real world applications. The study proposed by Athineos,2007 has explored how the specific factor performs when utterance lengths are considerably shortened. Essentially, the existing factor analysis systems which use different compensation methods have not shown any strong performance dissimilarities for short utterances. Most of the session variability compensation techniques used in factor analysis approaches are general. The performance of factor analysis based tactic drops severely when utterance lengths is decreased mainly less than ten seconds. Further efforts to explore the optimal compensation techniques are required. The work proposed by Sahidullah,2016 has derived a set of feature vectors using local covariance information. Further this study has also inspected how the local and global covariance information is related. Relative performance improvement 12.28% is obtained when the derived local features are combined with conventional cepstral features.

How to handle robustness of the authentication system while capturing the local uncertainty information from the speech segments is the unsolved problem. So there is a considerable scope for researchers to do further analysis in this direction. Some of the works in the literature suggests using complementary features besides using the conventional features for representing speaker specific information. The idea of Athineos,2007 is based on finding the relationship between the time domain envelope and autocorrelation of frequency domain envelopes of a speech segments. The constant Q transform based coefficients proposed by Todisco,2016 are substitute to traditional cepstral coefficients and it also resembles the human perception system more precisely. Combining CQCC and conventional cepstral coefficients extracted from the RSR2015 database have attained 60% reduction in the EER compared to cepstral features. This work could be extended further to analyze whether the fusion of CQCC features and time domain features would help in improving the system performance.

The aim of work suggested by Li,2015 is to analyze the complementary information captured by amplitude and phase features. The combined system has achieved 37% reduction in EER compared to that of cepstral features on NIST SRE 2008 database. The work suggested by Vijjayan, 2016 is based on the use of speaker specific phonemes to improve the discrimination ability of the classifier. The EER reduction of 4% is obtained when speaker phonetic information is utilized in classification task. Mary,2008 has experimentally demonstrated a new method of extracting prosodic features directly from speech segment. Syllable level sequence is acquired using vowel onset point as a reference without using any speech recognizer. Evaluation of proposed prosodic features on NIST SRE 2003 database has shown EER reduction of 2.5%. Many authors in literature have emphasized on exploring high level features . Thus high level features also have significance similar to low level features for speaker representation. From the studies which we have reviewed , we observe that most of the researchers have used spectral features and some of the studies have used combination of spectral features and long term features. However, exploring the best combination of these features which could uniquely characterize speaker in practical situation is the toughest challenge. The work made by Scheffer,2014 deals with the content mismatch issue of test and enrollment data at the statistical level.

The statistics of the enrollment data are transformed to that of the test data followed by predicting the speaker model related to test input. This system show performance boost by 50% for seen conditions .However, data from different content degrades the performance. This issue opens up research directions to analyze why

the content matching could not work in the practical senario. This encourage further study in this direction to design the decision tree taking into consideration more and more the realistic conditions. This could help in achieving better system performance. The study made by Park,2017 focuses on the issues related to large intra-speaker variability particularly for short utterances It is found that performance deviates to the large extent with seen and unseen intra-speaker variability which is not like humans. For that reason exploration of speaker characteristics insensitive to intra-speaker variability is the another big challenge which is yet to be resolved. We observe that works reviewed in this paper use different databases with different conditions for short utterances for evaluating the proposed systems. Therefore, direct comparison of their performance is not beneficial.

Another challenging problem in speaker verification is benchmarking the amount of speech data for obtaining SV systems with acceptable performance. This opens up scope for the researchers to study whether a speech utterance with the given duration is appropriate for enrollment or test and suggest some reliable performance measures in this direction. We also observe that the performance of the SV system is very much dependent on phonetic contents of the speech. This inspires further study to find out the features invariant to phonetic content of the speech data. Another problem which needs to be addressed is replay attack on the SV system by using the speech recordings of the claimed speaker to get access to the required service. With the importance of spoofing attack, development of systems in his direction has gained attention recently. Apart from the above discussed directions the deployment of SV system in presence of noisy conditions is of major concern .There are many methods for improving SU-SV system with constrained noisy conditions, however, very few for unconstrained noisy conditions. More advanced feature extraction approaches could be investigated relevant to short utterance

CONCLUSION

The issue of short utterance for speaker verification point of view has been considerably increased in recent studies. Many research efforts are carried out to handle this problem in different domains of speaker verification. In this paper, we have essentially done a brief review of feature extraction strategies that are used in the past and recent studies. In this review we have explored features extraction methods at different levels such as source level, vocal tract level and supra-segmental level. This review provides possible suggestions with which researchers could extend the existing work and solve the challenges in this area.

REFERENCES

A. K. Jain, A. Ross and S. Prabhakar, "An introduction to biometric recognition," 2004 IEEE Transactions on Circuits and Systems for Video Technology, vol. 14, no. 1, pp. 4-20, Jan. 2004, doi: 10.1109/TCSVT.2003.818349.

A. Kanagasundaram , R.Vogt , D. Dean, S. Sridharan and M. Mason, "i-vector based speaker recognition on short utterances,"2011 Proceedings of Interspeech August 2011, Florence, Italy.

A. Kounoudes, V. Kekatos and S. Mavromoustakos, "Voice Biometric Authentication for Enhancing Internet Service Security," 2006 2nd International Conference on Information & Communication Technologies, Damascus, 2006, pp. 1020-1025, doi: 10.1109/ICTTA.2006.1684514.

A. Matza ,Y. Bistritz, "Skew Gaussian mixture models for speaker recognition" 2011 IET Signal Process, Florence, Italy, vol.8,no.8,pp. 860–867, Aug.2011.

A.D. Mengistu, D. M. Alemayehu, "Text independent Amharic language speaker identification in noisy environments using speech processing techniques," 2017 Indonesian Journal of Electrical Engineering and Computer Science, vol. 5,no.1 ,2017,doi: 10.11591/ijeecs.v5.i1.pp109-114.

A. Viñals, A. Ortega, Miguel,E. Lleida, "An Analysis of the Short Utterance Problem for Speaker Characterization,"Applied Sciences,Switzerland2019,vol.9,No.18,pp.3697,Sep.2019,doi: 10.3390/app9183697

B. Bharathi and T. Nagarajan , "GMM and i-vector based speaker verification using speaker-specific-text for short utterances," 2013 IEEE International Conference of IEEE Region 10 (TENCON 2013), Xi'an, 2013, pp.1–4, doi: 10.1109/TENCON.2013.6718988.

C. Kim , R.M. Stern, "Power-normalized cepstral coefficients for robust speech recognition," 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, 2012, pp.4101–4104, doi: 10.1109/ICASSP.2012.6288820.

D. A. Reynolds, "Channel robust speaker verification via feature mapping," 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)., Hong Kong, 2003, pp. 53-56, doi: 10.1109/ICASSP.2003.1202292.

D. Hosseinzadeh , S. Krishnan, "On the use of complementary spectral features for speaker recognition," 2007 EURASIP Journal on Advances in Signal Processing, vol.1, Dec. 2007,doi: 10.1155/2008/258184.

D. Patil , S.R.M Pr.asanna, "A comparative study of explicit and implicit modeling of subsegmental speaker-specific excitation source information," 2013 Sadhana, vol. 38, no.4,pp. 591–620, 10.1007/s12046-013-0163-z.

D.A. Reynolds ,R.C. Rose , "Robust text-independent speaker identification using Gaussian mixture speaker models," 1995 IEEE Transactions on speech and audio processing, vol 3,no.1,pp.72–83, Jan. 1995,doi: 10.1109/89.365379.

D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker

Verification Using Adapted Gaussian Mixture Models", 2000 Digital Signal Processing, vol.10, pp.19-41, Jan.2000,doi: 10.1006/dspr.1999.0361.

E. Ambikairajah, "Emerging features for speaker recognition," 2007 6th International Conference on Information, Communications & Signal Processing, Singapore, 2007,pp. 1–7 , doi: 10.1109/ICICS.2007.4449889.

G. Doddington , " Speaker recognition based on idiolectal differences between speakers," 2001 Seventh European Conference on Speech Communication and Technology 2001, Aalborg, Denmark, pp. 2521–2524.

H.S. Jayanna ,S.R.M. Prasanna, " Multiple frame size and rate analysis for speaker recognition under limited data condition," 2009 IET Signal Processing,vol.3, pp. 189–204,May. 2009,doi: 10.1049/iet-spr.2008.0211.

J. Guo , G. Yeung , D. Muralidharan ,H. Arsikere ,A. Afshan and A. Alwan , " Speaker verification using short utterances with dnn-based estimation of subglottal acoustic features," 2016 Proceedings of Interspeech 2016, pp.2219–2222,doi: 10.21437/Interspeech.2016-282.

J. P. Campbell, "Speaker recognition: a tutorial," in Proceedings of the IEEE, vol. 85, no. 9, pp. 1437-1462, Sept. 1997, doi: 10.1109/5.628714.

J.Villalba and E. Lleida , "Preventing replay attacks on speaker verification systems," Proceedings of ICCST 2011,pp. 1–8.Oct.2011, doi: 10.1109/CCST.2011.6095943.

K. Veselý, M. Karafiát, F. Grézl, M. Janda and E. Egorova, "The language independent bottleneck features," 2012 IEEE Spoken Language Technology Workshop , Miami, FL, 2012, pp.336–341, doi: 10.1109/SLT.2012.6424246.

K. Vijayan, Reddy P, K.S.R Murty, "Significance of analytic phase of speech signals in speaker verification Speech Communication, vol.81, pp.54 – 71,2016.

K.Y. Leung, M.W. Mak, and S.Y. Kung , "Adapive articulatory feature-based conditional pronunciation modeling for speaker verification," 2006 Speech Communication,vol.48, No. 1,pp. 71 – 84,2006,doi: 10.1016/j.specom.2005.05.013.

L. Mary , B. Yegnanarayana , "Extraction and representation of prosodic features for language and speaker recognition," Speech Communication, vol. 50, no. 10, pp. 782 – 796,2008, doi: 10.1016/j.specom.2008.04.010.

L. Rodriguez -Linares and C. Garcia-Mateo, "A novel technique for the combination of utterance and speaker verification systems in a text-dependent speaker verification task," 1998 Fifth International Conference on Spoken Language Processing , Sydney, Australia vol. 2,pp. 213–216,Nov.-Dec 1998.

M. Athineos and D.P.S. Ellis, " Autoregressive modeling of temporal envelopes," 2007 IEEE Transactions on Signal Processing,vol.55, no. 11, pp .5237–5245, Nov. 2007, doi: 10.1109/TSP.2007.898783.

M. Sahidullah , M. Kinnunen, "Local spectral variability features for speaker verification," 2016 Digital Signal Processing, vol.50 , no.1, pp .1–11 ,doi: 10.1016/j.dsp.2015.10.011.

M. Todisco, H. Delgado, N. Evans," Articulation rate filtering of cqcc features for automatic speaker verification" 2016 Proceedings of Interspeech, San Francisco, USA pp.3628–3632,doi: 10.21437/Interspeech.2016-1140.

M.J. Alam ,P. Kenny and T. Stafylakis, "Combining amplitude and phase-based features for speaker verification with short duration utterances," 2015 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, Sept. 2015, pp. 249–253.

N. Dehak, P. Dumouchel, P. Kenny, "Modeling prosodic features with joint factor analysis for speaker verification," 2007 IEEE Transactions on Audio, Speech, and Language Processing, vol.15,No.7,pp. 2095–2103, Sept. 2007, doi: 10.1109/TASL.2007.902758.

N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," 2011 IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 788-798, May 2011, doi: 10.1109/TASL.2010.2064307.

N. Dehak, P. Kenny andP. Dumouchel, "Comparison between factor analysis and GMM support vector machines for speaker verification," Proceedings of Speaker and Language Recognition Workshop 2008,IEEE-Odyssey, Stellenbosch, South Africa.

N. Fatima and T.F. Zheng, "Vowel-category based short utterance speaker recognition," 2012 Proceedings of International Conference on Systems and Informatics,Yantai,2012,pp. 1774–1778, doi: 10.1109/ICSAI.2012.6223387.

N. Fatima and T. F. Zheng, "Short Utterance Speaker Recognition A research Agenda," 2012 International Conference on Systems and Informatics (ICSAI2012), Yantai, 2012, pp. 1746-1750, doi: 10.1109/ICSAI.2012.6223381.

N. Scheffe and Y. Lei, "Content matching for short duration speaker recognition," 2014 Proceedings of Interspeech , pp.1317–1321,Sep.2014,

P. Kenny, G. Boulianne, P. Ouellet and P. Dumouchel, "Joint Factor Analysis Versus Eigenchannels in Speaker Recognition," 2007 IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 4, pp. 1435-1447, May 2007, doi: 10.1109/TASL.2006.881693.

P. Kenny, Joint factor analysis of speaker and session

variability: theory and algorithms. Technical Report, Jan.2006.

S.J .Park, G. Yeung, N. Vesselinova , Kreiman, Keating P,.Alwan A, "Towards understanding speaker discrimination abilities in humans and machines for text-independent short utterances of different speech styles," 2018 Journal of the Acoustical Society of America,vol.144,No. 1,pp.375–386,2018,July. 2018, doi: 10.1121/1.5045323.

S.J.D. Prince and J.H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," 2007 IEEE 11th International Conference on Computer Vision, Rio de Janeiro, 2007, pp. 1-8, doi: 10.1109/ICCV.2007.4409052.

S.J.Park , G. Yeung ,P.A. KreimanA , P.A. Keating and A. Alwan, "Using Voice Quality Features to Improve Short-Utterance, Text-Independent Speaker Verification Systems," Proceedings of Interspeech 2017, pp.1522–1526,doi: 10.21437/Interspeech.2017-157.

S.O. Sadjad, J.H. Hansen, "Mean hilbert envelope coefficients for robust speaker and language identification," 2015 Speech Communication, vol.72,pp. 138 – 148,doi: 10.1016/j.specom.2015.04.005.

S.Paulose, D. Mathew , A. Thomas ," Performance Evaluation of Different Modeling Methods and Classifiers with MFCC and IHC Features for Speaker Recognition," 2017 7th International Conference on Advances in Computing & Communications, ICACC-2017,Cochin,vol.115,pp.55-62,doi: 10.1016/j.procs.2017.09.076.

S.R.M. Prasanna , C.S. Gupta, B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech" 2006 Speech Communication,vol.48,no.10,pp. 1243–1261,2006, ,doi:1016/j.specom.2006.06.002.

T. Fu ,Y. Qian ,Y. Liu and K. Yu, "Tandem deep features for text dependent speaker verification," Proceedings of Interspeech,2014,pp.1327–1331.

T. N. Sainath, B. Kingsbury and B. Ramabhadran, "Auto-encoder bottleneck features using deep belief networks," 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, 2012, pp. 4153-4156, doi: 10.1109/ICASSP.2012.6288833.

W. Andrews, W. Kohler and J. Campbell J, "Gender-Dependent Phonetic Refraction for Speaker recognition," 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, 2002, vol.1, pp.149-152, doi: 10.1109/ICASSP.2002.5743676.

W.M. Campbell, J. Campbell, D.A. Reynolds ,E. Singer, P. Torres- Carrasquillo , "Support vector machines for speaker and language recognition," 2006 Computer Speech & Language,vol. 20,no. 3,pp. 210–229,2006,doi: 10.1016/j.csl.2005.06.003.

Y Li, W.Q. Zhang, J. Liu , "Multi-resolution time frequency feature and complementary combination for short utterance speaker recognition," 2015 Multimedia Tools and Applications, vol.74,no.3,pp. 937–953,doi: 10.1007/s11042-013-1705-4.

Y. Lei, N. Scheffer, L. Ferrer and M. McLaren, "A novel scheme for speaker recognition using a phonetically aware deep neural network," 2014 IEEE International conference on Acoustics, Speech, and Signal Processing (ICASSP), Florence, 2014, pp. 1695-1699, doi: 10.1109/ICASSP.2014.6853887.

Y.Chen and Z.M. Tang , "The speaker recognition of noisy short utterance," 2013 Proceedings of Intelligence Science and Big Data Engineering. Lecture Notes in Computer Science, vol 8261. Springer, Berlin, Heidelberg,vol.8261,pp. 666–671,doi:10.1007/978-3-642-42057-3_84.