Expert Systems With Applications 90 (2017) 250–271

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Review

Speaker identification features extraction methods: A systematic review

Sreenivas Sremath Tirumala[a,1], Seyed Reza Shahamiri[a,*], Abhimanyu Singh Garhwal[a,1], Ruili Wang[b,2]

[a] Faculty of Business and Information Technology, Manukau Institute of Technology, Auckland, New Zealand
[b] Computer Science and Information Technology, Institute of Natural and Mathematical Sciences (INMS), Massey University, Auckland, New Zealand

ARTICLE INFO

Article history:
Received 12 May 2017
Revised 4 August 2017
Accepted 6 August 2017
Available online 16 August 2017

Keywords:
Feature extraction
Kitchenham systematic review
MFCC
Speaker identification
Speaker recognition

ABSTRACT

Speaker Identification (SI) is the process of identifying the speaker from a given utterance by comparing the voice biometrics of the utterance with those utterance models stored beforehand. SI technologies are taken a new direction due to the advances in artificial intelligence and have been used widely in various domains. Feature extraction is one of the most important aspects of SI, which significantly influences the SI process and performance. This systematic review is conducted to identify, compare, and analyze various feature extraction approaches, methods, and algorithms of SI to provide a reference on feature extraction approaches for SI applications and future studies. The review was conducted according to Kitchenham systematic review methodology and guidelines, and provides an in-depth analysis on proposals and implementations of SI feature extraction methods discussed in the literature between year 2011 and 2106. Three research questions were determined and an initial set of 535 publications were identified to answer the questions. After applying exclusion criteria 160 related publications were shortlisted and reviewed in this paper; these papers were considered to answer the research questions. Results indicate that pure Mel-Frequency Cepstral Coefficients (MFCCs) based feature extraction approaches have been used more than any other approach. Furthermore, other MFCC variations, such as MFCC fusion and cleansing approaches, are proven to be very popular as well. This study identified that the current SI research trend is to develop a robust universal SI framework to address the important problems of SI such as adaptability, complexity, multi-lingual recognition, and noise robustness. The results presented in this research are based on past publications, citations, and number of implementations with citations being most relevant. This paper also presents the general process of SI.

1. Introduction

Speech is a universal form of communication. Speaker Recognition (SR) is the process of identifying the speaker according to the vocal features of the given speech. This is different to speech recognition where the identification process is confined to the content rather than speaker. The process of SR is based on identifying and extracting unique characteristics of the speaker's speech. The characteristics of voices of the person is also known as voice biometrics.

A SR system is used to identify and distinguish speakers and extract unique characteristics that may be used for user verification or authentication. Speaker Identification (SI) is known as the process of identifying the speaker from a given utterance by comparing voice biometrics of the given sample of the speaker. When voice is used for authorization, it is termed as Speaker Verification. The key application area of SR is security and forensic science. SR systems are also used as a replacement for password and other user authentication processes (voiced password). Forensic science applies SR to compare the voice samples of the person claimed to be with other evidences obtained like telephone conversation or other recorded evidence. This process is also referred as speaker detection. The most important aspect of using SI systems is for automating processes like directing clients' mails to the right

* Corresponding author at: MIT Manukau, Cnr of Manukau Station Rd Davies Ave, Private Bag 94006, Manukau 2241, New Zealand.
E-mail addresses: ssremath@aut.ac.nz (S.S. Tirumala), admin@rezanet.com, rshahamiri@gmail.com, rshahamiri@yahoo.com (S.R. Shahamiri), abhimanyu garhwal@gmail.com (A.S. Garhwal), Ruili.wang@massey.ac.nz (R. Wang).
[1] MIT Manukau, Cnr of Manukau Station Rd Davies Ave, Manukau, Private Bag 94006, Manukau 2241, New Zealand.
[2] Room 3.10, IIMS Building, Albany Campus, Massey University, Albany, Auckland, New Zealand.

http://dx.doi.org/10.1016/j.eswa.2017.08.015
0957-4174/

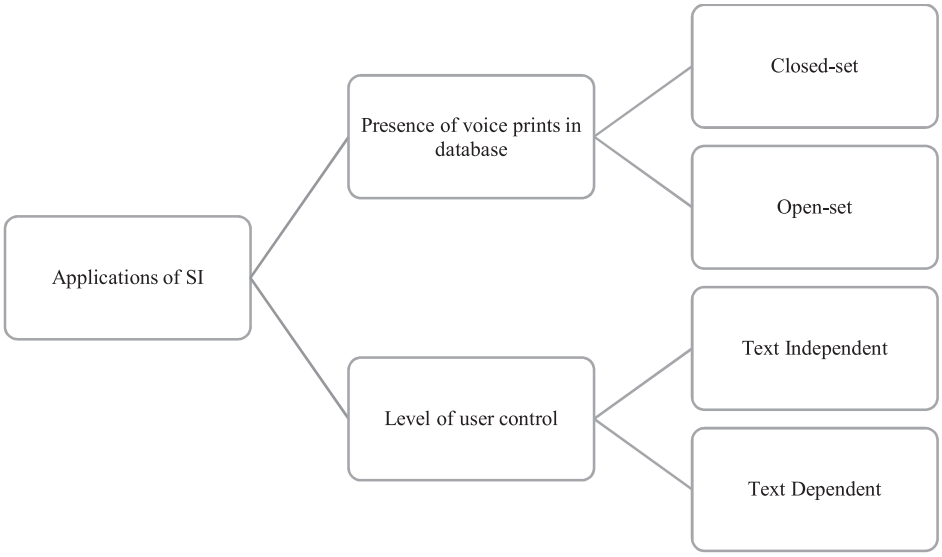

Fig. 1. Speaker identification applications classification.

mailbox, recognizing talkers in discussion, cautioning discourse acknowledgment frameworks of speaker changes, checking if a client is enlisted in the framework as of, and so on. These SI systems may work without the knowledge of a client's voice sample since they rely only on identifying an input speaker from the existing database of speakers.

Our systematic review is confined to SI as one of the primary types of SR systems (Reynolds, 2002). Feature extraction is one of the important SI aspects that significantly influences the quality of SI. In particular, the selection of proper feature extraction approaches plays a vital role since the identification is carried out by comparing unique characteristic features of a voice input. Therefore, the aim of this article is to carry out a systematic literature review on various feature extraction approaches of SI in order to:

(1) Identify significant feature extraction approaches in the last six years,
(2) Present a systematic review on the research of feature extraction approaches for SI,
(3) Classify various feature extraction approaches and provide recommendations based on the research.

The applications of SI can be classified in two types as presented in Fig. 1. The first type depends on the presence of voice prints in the database, which is further classified into two categories namely closed-set and open-set. In closed-set, the test speaker input is compared with the existing speakers' voice prints in the database and the nearest match is found (Dutta, Patgiri, Sarma, & Sarma, 2015). Hence, a closed-set SI guarantees a result although it may not be the exact speaker. On the other hand, in an open-set SI the input speaker voice print is compared with the database for 'exact match'; the input is rejected if the match is not found (Reynolds, 2002).

The second type of SI applications is based on the level of user control, which is also known as speaker verification process. This SI category is also further classified into two categories: text dependent and text independent. In text dependent the speaker must utter the same phrases or words that are previously used for training (Islam & Rahman, 2009; Kekre, Athawale, & Desai, 2011), while in the last category the input voice print content may not exist in the training set (Boujelbene, Mezghanni, & Ellouze, 2009; Revathi & Venkataramani, 2009; Verma, 2011).

The systematic review is carried out using Systematic Literature Review (SLR) methodology proposed by Kitchenham and Charters (2007) which is detailed in the methodology section. In this review, we presented various feature extraction approaches that were used in speaker identification processes and provide a systematic review on the research of these approaches. It is noteworthy to observe the key components of SI systems (which are detailed in the next sections) like parametrization (i.e. feature extraction), speaker modelling, pattern matching and scoring method that are core components for SR as well. This systematic review explained all SI components but more emphasis was put on feature extraction.

Since SR can be considered as a pattern recognition problem, various Artificial Intelligence (AI) approaches are used for SR systems (Rajesh et al., 2012). Deep Learning, which attained state of the art results for complex pattern recognition problems, has also been implemented for SR systems (Ghahabi & Hernando, 2014; McLaren, Lei, & Ferrer, 2015; Richardson, Reynolds, & Dehak, 2015b). Recent deep learning implementations for SR highlights the complexity involved in SR which requires special attention compared with traditional pattern recognition problems in general, and speech recognition problems in particular (Richardson, Reynolds, & Dehak, 2015a).

Existing review works and surveys on speaker recognition can be broadly categorized into three categories. The first category is the comprehensive surveys on SR that review the literature on generic SR processes and different SR categories (the SR categories are explained in the next section). There are numerous works in this aspect, such as El Ayadi, Kamel, and Karray (2011); Lawson, et al. (2011); Saquib, Salam, Nair, Pandey, & Joshi, 2010a). The second category mostly focuses on the types of statistical and machine learning approaches used as SR classifiers, for example Farrell, Mammone, and Assaleh (1994); Larcher, Lee, Ma, and Li (2014); Lippmann (1989). This category of SR reviews mostly falls under classification and machine learning research where there is considerable amount of literature available. In terms of speaker identification, the only work that specifically discussed SI and its processes is a brief survey presented by Sidorov, Schmitt, Zablotskiy, and Minker (2013) in which few generic SI methods were explained and compared.

The third category of speaker recognition surveys deals with feature extraction approaches in SR. One of the most recent SR feature extraction surveys is Dişken, Tüfekçi, Saribulut, and Çevik (2017) that concentrated on methods for extracting robust speaker

specific features based on noise profiles, emotion and channel mismatch. Another example is Rao and Sarkar (2014) that presented simplified explanation on model and feature based speaker verification systems. A short review by Chavan and Chougule (2012) was also provided that briefly defined and explained features in respect to speaker recognition. Another recent work that highlighted the importance of using deep learning approaches for feature extraction in speaker identification is Tirumala and Shahamiri (2016). Other examples are a review on SR analysis, modelling and feature extraction presented by Jayanna and Prasanna (2009), and a survey on evaluating SR acoustic features based on experimental evaluations (Lawson et al., 2011).

None of the existing SR review works carried out using a systematic review methodology. Furthermore, these works were not confined to a particular period and produced an overall progress report of SR research instead. Finally, they did not present a detailed analysis of speaker identification process and highlight the importance of feature extraction in SI. Thus, there is a gap in the literature in providing a systematic reference for SI feature extractions approaches. This paper tries to address this gap by providing a systematic review presenting a detailed overview of the majority of statistical and machine learning recent features extractions approaches. We also presented the speaker identification process in detail. In particular, this paper collected the related information and discussed SI technologies reported in the literature from 2011 to 2016 with special attention to SI feature extraction methods. It is pertinent to note that scoring methods are beyond the scope of this study. The key contribution of this paper is in collating all related SI implementations in one place, which will serve as a reference for SI researchers. Furthermore, this paper can be used to suggest criteria for selecting a particular feature extraction model for implementing SI systems.

2. Background

This section presents speaker recognition classifications and then describes the process of speaker identification.

2.1. Speaker recognition categories

In a research perspective, SR can be categorized according to the action performed, or on the basis of the research field, as shown in Fig. 2. In particular, action-based SR areas are:

(1) *Speaker Identification (SI)* to identify an unknown user based on her voice prints (Daoudi, Jourani, Andre, & Aboutajdine, 2011; Wu & Lin, 2009). It is the process to compare one user voice profile against many profiles and find the best or exact match.
(2) *Speaker Verification (or Authentication)* is the process of verifying the identity of a user by using her voice prints when the speaker claims to be a specific user (Jiang, Gao, & Han, 2009). It is a one to one match.
(3) *Speaker Diarization* is identifying a person's voice from the given population, and when the speaker speaks (Anguera et al., 2012; Poignant, Besacier, & Quénot, 2015). SI is different from speaker diarization: in SI the input is typically only one-user voice and the objective is to match the speech features to a speaker profile from the data source. Nevertheless, in speaker diarization, a mixture of utterances from various users is given to the system while the system's objective is to identify a specific user's speech and determine when she speaks.
(4) *Speaker De-Identification* is used to maintain anonymity of the users. It is commonly employed to hide the identity of the users where their identity must be hidden while maintaining the acoustic information from speech (Jin, Toth, Schultz, & Black, 2009; Justin et al., 2015; Pobar & Ipsic, 2014).
(5) *Voice Activity Detection (VAD)* is the process of determining the existence of human speech (Haigh & Mason, 1993; Ram, Segura, Ben, De La Torre, & Rubio, 2004).

Research-based SR areas are as follows:

(1) *Speaker Modelling* is the process of identifying and associating a unique identifier to the voice prints of an individual speaker in order to differentiate from other speakers presented in the database (Beigi, 2011).
(2) *Speech Parameterization* is the process of calculating a set of parameters from a small portion of speaker's voice prints that describes the properties of the speaker or speech signal (Ganchev, 2011).
(3) *Pattern Matching and Scoring Methods* are used to compare patterns presented in the input speaker's voice prints with the patterns extracted from various speakers in order to match unique characteristic features. Then, each similarity match is given a score that determines the accuracy of speaker identification.

The SI process and phases are described in the following.

2.2. Speaker identification process

In general, a SI system goes through two primary phases: a training phase that is also called enrolment, and a matching phase where the enrolment is verified for a match. A typical block diagram for both phases of SI is given in Fig. 3 though some SI techniques may bypass certain steps.

The enrolment phase starts with receiving the modelling speech input signals and data pre-processing and normalization. The next step is feature extraction that is providing the speech signal parameters in such a way that is understandable by the system. The extracted features may need normalization too before the training process commences. The training process may involve both offline training (training algorithm, background modelling, model adaptation) followed by online training (model adaptation). The results of the training phase are speaker models that are stored to be used in the next phase. Section 2.3 explains this phase in more details.

The objective of the matching phase is to match a speech signal obtained from the speaker to be identified (i.e. the test speaker) against the speaker models stored during the enrollment phase in order to identify the speaker uttering the speech. Similar to the first phase the input signal requires to be pre-processed, normalized, and its features need to be extracted. Next, the test speaker features are compared with the trained speaker models looking for a match. This is followed by calculation of similarity score and normalizing it. Section 2.4 provides more information about the matching phase.

2.3. SI phase 1: speaker enrolment or training phase

This phase is initiated by speech parametrization in which the speech input is pre-processed and normalized before extracting the features. The extracted speech features may also need normalization before creating and storing the speaker voice prints or models for training. The process of speech parametrization is:

(1) Inputting the speech signal: the following parameters of the input signals are required to be considered:
  – *The source of the speech:* to determine whether the speech signal is from a live subject or it is a recorded speech.
  – *Language:* different languages may highlight different types of speech features that influence the performance of the SI. Although most of the SI systems in the literature are designed for English, the literature reports non-English based SI systems as well. For example: Japanese (Kawakami, Wang,

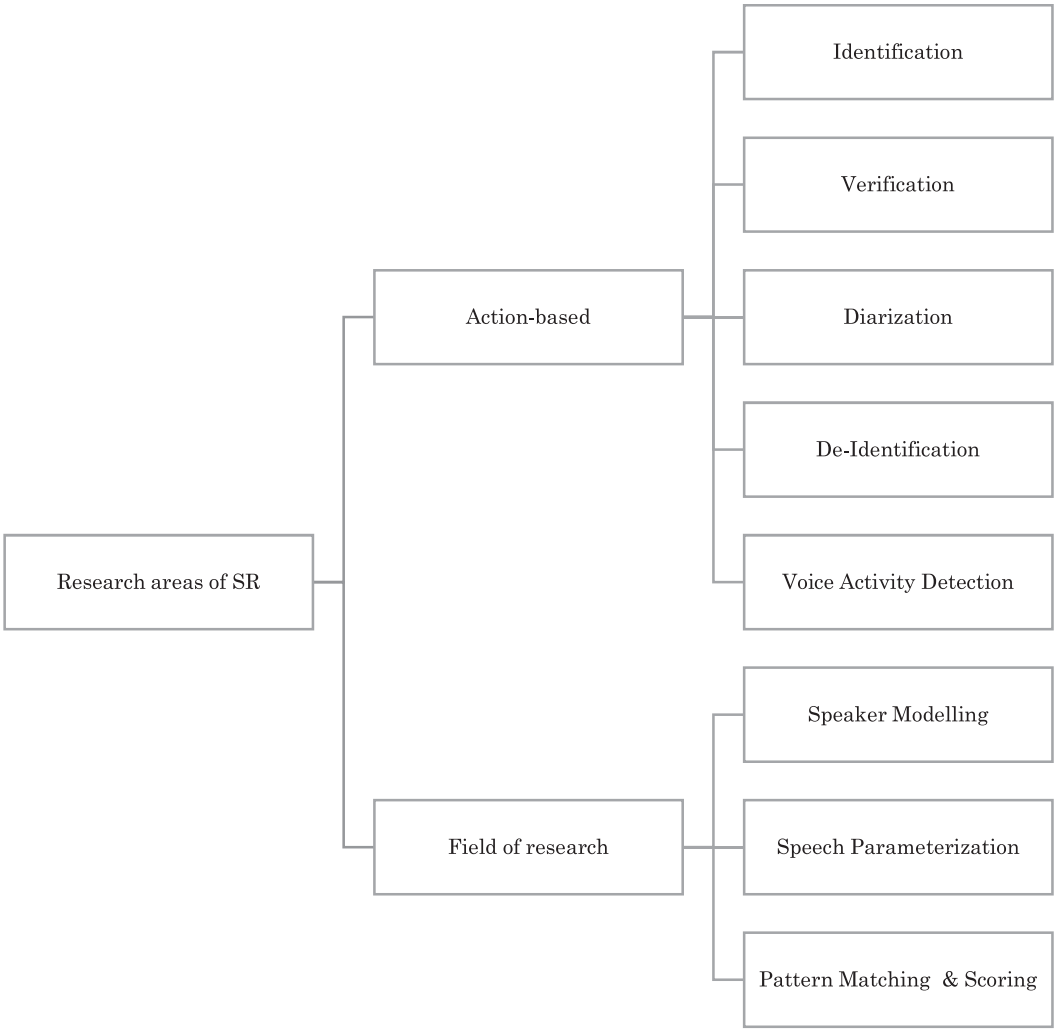

Fig. 2. Main areas of research for speaker recognition.

Kai, & Nakagawa, 2014), Thai (Tanprasert & Achariyakulporn, 2000), Spanish (Luengo et al., 2008). Additionally there are some notable Indian languages like Assamese (Sarma & Sarma, 2013a), Marathi (Jawarkar, Holambe, & Basu, 2012), and Hindi (Jawarkar, Holambe, & Basu, 2015). There are also few multilingual speaker identification approaches like (Nagaraja & Jayanna, 2013).

- *Speech capturing device:* different types of capturing devices record the speech differently since they have different types of sensors, level of capturing capabilities, and sensitivities. For example, some types of microphones are designed to capture speech signals for a particular environment. There are also microphones that are designed for a particular purpose like portable microphones, noise cancelation microphones, computer microphones, and microphones embedded in telephones or smartphones.
- *Environmental Noise:* different noise profiles like background noise, environmental noise, and room echoes may disrupt the input signal and significantly affect the performance of speech processing systems. Microphone sensitivity is also influenced by noise; for instance, in quiet conditions, a highly sensitive microphone may capture not only the original sound of the speaker but also the reverberation signals (Zhang et al., 2015). Noise robustness methods or smart room environments may be used to reduce the effects of noise (Busso et al., 2005; Shahamiri & Salim, 2014c).
- *Speech Variability:* the manner of a speaker such as rate, volume, sickness, age, emotions, time of the day (morning vs. evening voice for example), etc. may modify the speech features too. An example is the study conducted to investigate the effects of six emotional states (neutral, happy, sad, angry, disgust and fear) in SI systems by Sahidullah, Chakroborty, and Saha (2011). Mood identification of a speaker was also considered in Ahmed, Kenkeremath, and Stankovic (2015).

(2) Pre-processing of the speech: this process deals with any hindrances or glitches that may affect feature extraction. It mainly tries to remove noises and silence gaps. This is important because noises and silence gaps in the inputs possess highly non-stationary characteristics that can cause false identification (Farhood & Abdulghafour, 2010; Keerio, Mitra, Birch, Young, & Chatwin, 2009).

(3) Normalization: it helps to remove any variations like intersession variability and variability over time that may cause the speech features to fluctuate at the cost of some feature loss. This intersession variability is due to the changes in recording environment, transmission circumstances, background noise and variations in speaker voices. Identical speaker utterances cannot be repeated in a similar manner for each and every trial

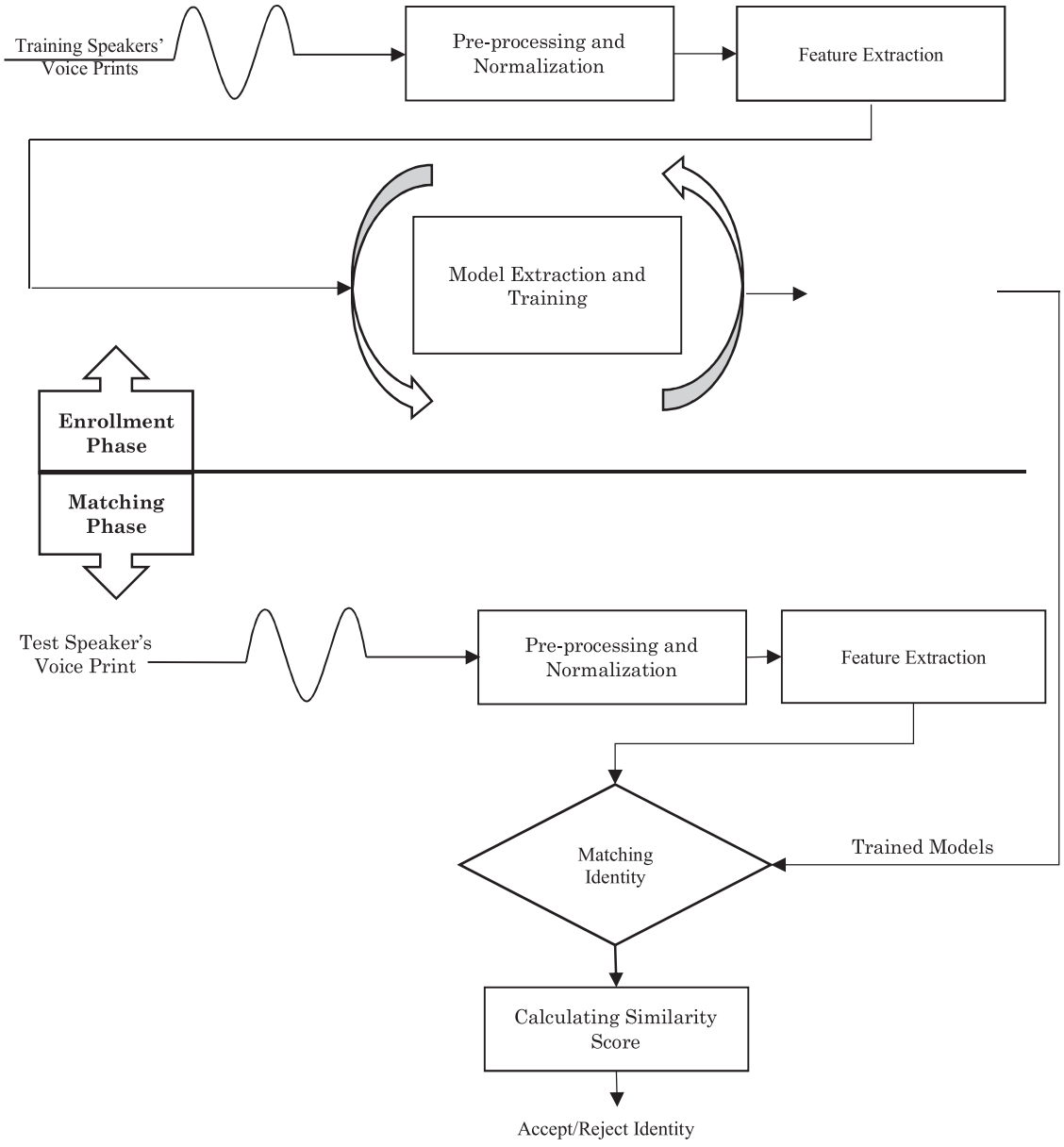

Fig. 3. Phases of Speaker Identification.

i.e. the utterance recorded in one session is highly correlated to recordings in a different session. Fig. 4 depicts the factors responsible for intersession variability.

A common practice for normalization is to use filter banks. There are two types of normalization techniques namely parameter based and distance (or similarity) based. The former type of normalization is proved efficient to decrease the effects of long spectral variation and linear channels (Atal, 1974; Furui, 1981). Text-dependent speaker recognition systems that have sufficiently long utterances apply a process called spectral equalization (aka blind equalization), which it is the process of removing interferences or noise through a non-linear phrase response signal. In this process, the total average value of the cepstral coefficients of the total utterances are subtracted from the averaged values of cepstral coefficients of an individual frame. Nevertheless, the side effect of this process is it removes some speaker specific and text-dependent features; hence it is not useful for SR systems with short utterances.

On the other hand, the distance (or similarity) domain normalization technique approximates optimal Bayes scoring. To put it differently, given the utterance's observed measurements, it is the (posteriori) ratio of these two conditional probabilities (Perner, 2010). Distance similarity domain technique is quite useful accommodating the variability factor in the speech signal thresholds for each speaker. Since this technique uses posteriori ratio, the noise signals can be easily differentiated. As an illustration HMM-based SR for noisy conditions and Parallel Model Combination (PMC) were successfully applied with the application of distance/similarity domain normalization approach (Gales & Young, 1992). Posteriori ratio can be implemented when the test speaker is presented in the repository since the calculation are performed for all the speakers in the database including the speaker. Nevertheless, this approach is not practical in case the test speaker voice prints are not considered during the training phase.

(4) Feature extraction: it is the process of presenting an acoustic signal as specific acoustic features. The features are selected to

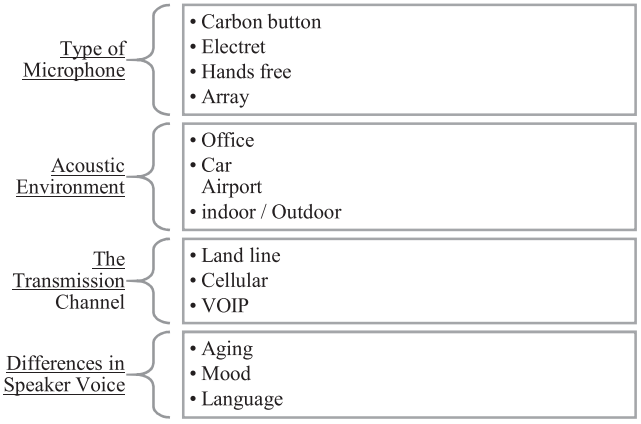

Fig. 4. Factors responsible for intersession variability.

best present the acoustic characteristic of the signals for different types of speech processing systems. It is discussed in details in the next section.

Speaker modelling is the next step of the SI training phase in which speaker models are trained using the extracted acoustic features of the speakers. The efficiency of speaker models is reflected in identifying the speaker accurately with the objective of minimizing the error rate. There are three types of modelling techniques: classical approaches, parametric approaches based on a training paradigm, and hybrid techniques that apply machine learning techniques. A hierarchy for various speaker modelling approaches is presented in Fig. 5.

The classical approach has two types of models; template-based models which are based on vector quantization, dynamic time wrapping, or histogram models (Ezequiel, 2014), or stochastic models based on Gaussian Mixture Model (GMM) or Hidden Markov Model (HMM) (Campbell, 1997).

The training paradigm models (parametric approaches) can be Generative (like GMMs or Vector Quantization) or Discriminative models (generally using machine learning techniques like SVMs and ANNs). The hybrid approaches are a combination of the above models such as GMM-HMM, ANN-HMM, etc. Further, some approaches like Vector Quantization and GMMs can be classified in both classical and training paradigms.

The parametric approaches fit some distribution to the training data by searching for the parameters of the distribution that maximizes the required criterion. The non-parametric approaches, on the other hand, make minimal assumptions about the distribution of the features.

The training process can be either offline or online. Offline training models require a fixed repository of speakers. It is necessary for training discriminative models to use the speakers' data from the repository as negative samples for recognition. In addition, offline training approaches are used in Universal Background Modelling (UBM) for unique adaptation of speakers based on the speakers' feature vectors. Thus, offline training requires all the known speaker models for training (Sreenivasa Rao & Sarkar, 2014).

UBM is used in online training. However, the speaker models are adopted 'online' from the training data of the speakers in real time. In particular, the models are applied in real-time which makes it more robust for identifying unknown speakers (i.e. the models that are not presented in training). This adaptation further helps to expedite the second phase of SI that is explained in the next section.

2.4. SI phase 2: identification or matching phase

The identification phase starts with speech parametrization which was detailed in the previous section. This is followed by the identity matching step where features extracted from an unknown speaker utterances are given to the system in order to identify the speaker. Similarity scores (i.e. likelihood) are also produced comparing the given input utterance with any speaker model stored in the system. The pattern matching is probabilistic in stochastic models where results are calculated in the form of conditional probabilities; on the other hand, template models employ deterministic approaches.

The matching part of this phase usually applies pattern matching algorithms such as those shown in Fig. 5. It is responsible for identification of the speaker by matching the trained speaker models using the extracted features from the unknown speaker utterance. To determine the best match, the identification process compares the utterance against multiple speaker models, or voice prints.

3. Research methodology

Systematic literature review (SLR) guidelines proposed by Kitchenham and Charters (2007) and Kitchenham et al. (2010) were considered to conduct this research. This SLR aims in understanding the contribution of the earlier works and identifying research gaps. The stages of the SLR approach adopted from Champiri, Shahamiri, and Salim (2015) is presented in Fig. 6 and the process is discussed below.

3.1. Planning the review

Throughout planning we defined the objectives of the proposed SLR and performed the required assessments. The planning process was:

(1) *Identification of need of SLR:* with the process of planning we have identified that there is no recent SLR presented in the area of SI that emphasizes on speaker modelling and speech parametrization. Although there are few research review publications on speech and speaker recognition generic topics such as (Furui, 2005, 2009; Saquib, Salam, Nair, Pandey, & Joshi, 2010b), none of them presented a systematic review and discussed speaker identification specifically. An overview of the methods that can be applied to improve SI accuracy and robustness where speakers'' distinguishable data are missing is provided by Togneri and Pullella (2011). Similarly, there was an overview paper published in 2010 reviewing the shift from SI vector models to super-vector paradigms between 1980 and 2010 (Kinnunen & Li, 2010). Nevertheless, this overview was not conducted systematically. Moreover, there has been several new methods introduced recently in the context on SI feature extraction that need to be systematically reviewed. This paper concentrates on filling in this gap by identifying and summarizing feature extraction evolution in SI specifically within the period of 2011 to the end of 2016. This systematic review also provides a comprehensive research reference for SI researchers.
(2) *Formulating the questions:* we formulated the following questions for this review:
   - What are the criteria for optimal features and how feature parameter appropriateness is decided for the feature extraction process in SI?
   - What are the feature extraction approaches and algorithms used in SI process?
   - What is the most popular and successful feature extraction approach in the last six years?

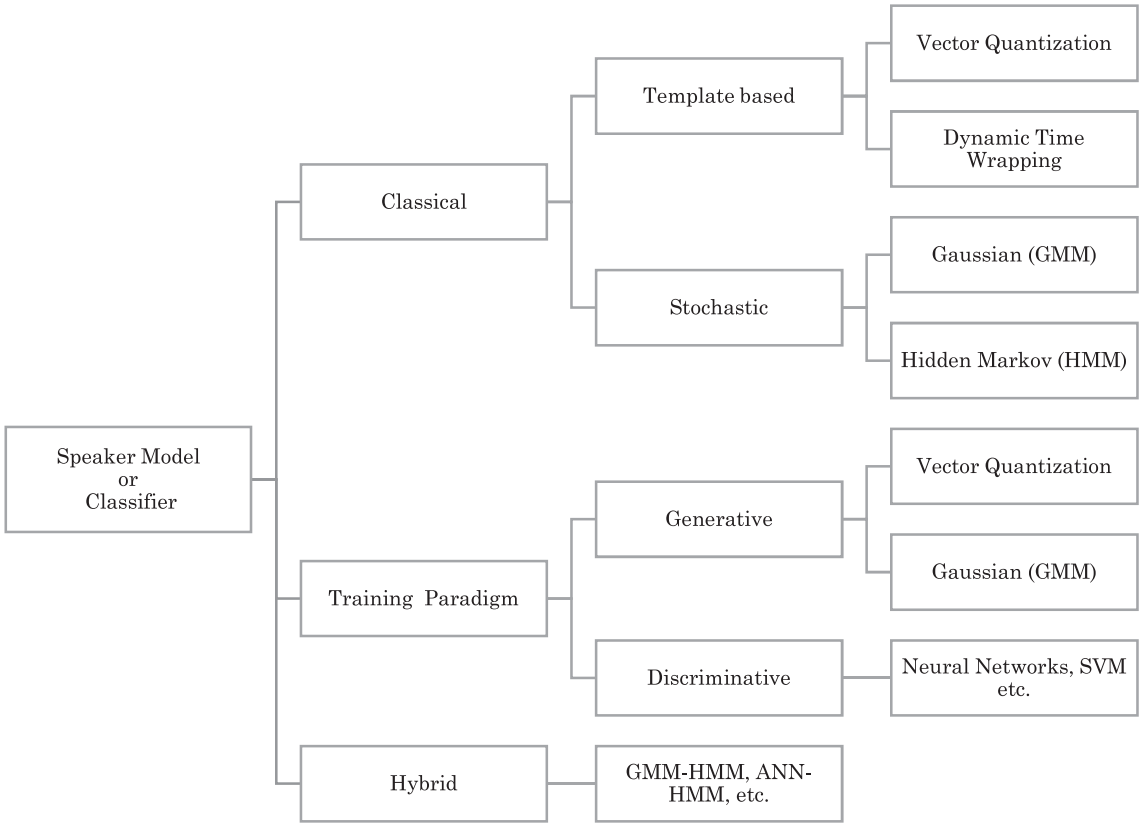

Fig. 5. Speaker modelling approaches used in SI.

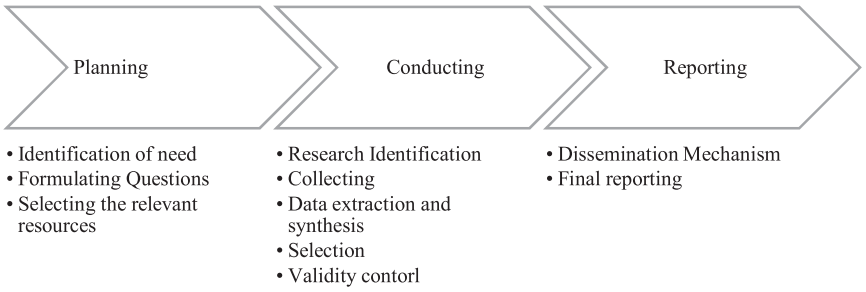

Fig. 6. SLR steps and activities.

(3) *Selection of the relevant resources:* the investigation process was carried out according to the SLR guidelines; the search process was conducted for a fixed start and end dates specifying month and year as recommended in Stapic, Lo, Cabot, de Marcos Ortega, and Strahonja (2012). Popular digital resources like IEEE Xplore Digital Library, ScienceDirect, ACM Digital Library, Google Scholar, DBLP, and Springer Verlag were considered during the search. As this research is focused on the last six years (at the time of commencing this study), the date range for the search process was limited to January 2011 and December 2016. Nonetheless, the time limitation was not applied when resources were required to explain general SI processes or techniques.

3.2. Conducting the review

The process of conducting the review is explained in the following:

(1) *Research identification:* initial search using speaker recognition as the keyword without any filters resulted in 1710 papers from IEEE Xplore, 34,045 from ScienceDirect, 6678 from ACM digital library, and 36,676 from Springer Verlag. Resources with combined results from multiple sources like Google Scholar and DBLP resulted in 1160,000 and 764 entries respectively. Changing the keyword to speaker identification and applying the time limitation filter a total of 190 relevant papers were identified. This step is further explained in the "selection" step.

(2) *Strategy for collecting the primary studies:* the significance of the 190 papers identified in the last step were studied by reading each paper's abstract, methodology, discussion, and results sections thoroughly. The following criteria were applied for classification and papers that did not meet them were excluded:
   - The paper discussed speaker identification for academic or scientific digital libraries.
   - The paper discussed speaker identification in order to provide recommendations for books or articles.
   - The paper discussed speaker identification in order to provide recommendations of SI techniques for academic or scientific digital libraries, or scientific document recommendations.

- The experiment data set is relevant to the SI process which includes feature extraction as a part of it.
- The speaker identification method discussed in the paper was created for academic and scientific audience.
- The speaker identification method discussed in the paper was created for any practical use in biometric applications.

Additional exclusion criteria that were considered in this paper were:

- If the document discussed speaker identification in order to perform collection acquisition, selecting the materials, cataloguing and organizing, and disseminating the information.
- No peer review assessment was performed or reported.
- The SI processes that were a part of speaker diarisation, transmission channel like Skype, telephone etc. and language identification process.
- If a speaker identification paper only provides a comparison method for the speaker diarization in different scenarios.
- Full text of the paper is not available.
- Any paper consisting any other identification process of the speakers.

(3) *Extracting data and synthesizing:* SLR rules for extracting and synthesizing data were followed for each paper as explained by Francc, Gouveia, Santos, Santana, and da Silva (2011). The papers were shortlisted based on answers to the research question(s) provided by the research papers. These results were recorded on a different results structure by distinguishing the subjects from the discoveries reported in each acknowledged paper. For our situation, these recognized topics and criteria gave us the classifications reported in our findings and results segment.

The processor of data extraction can be defined as the amount of classification provided throughout data extraction, and also the amount provided in the data synthesis step. SLR with Kitchenham's recommendations does not provide a detailed and clear way about the process of data extraction. Hence, we selected on trivial data extraction bringing about a record of quotes that were just insignificantly reworded; in the synthesis step such quotes were early classified. In this section, we exhibit frequencies of the quantity of times every subject is recognized in various sources in which every event was given a similar weight. Such frequencies only reflect how frequently a given issue is distinguished in various papers. Nevertheless, they cannot determine how vital it might be.

(4) *Selection:* during document retrieval process, initially we retrieved 535 publications by applying the keyword "Speaker Identification" but without the time filter. Applying the time limitation, the number was reduced to 190. Two papers were rejected on the basis of the language other than English. By analyzing the title and abstract another 21 papers were filtered leaving out 167. Then we looked at the availability, exclusion criteria discussed before, and various other consequences, and the final list of 159 papers were shortlisted. The detailed statistics of the selected papers at various steps is shown in Table 1.

(5) *Validity control:* from the total of 190 papers retrieved, 30% (38 papers) were randomly selected and verified with another author from the same field of research in order to affirm the relevance of the 190 papers. This process resulted in 84% of the papers (32 out of 38) were found to be relevant. Next, we composed a list from the references of the 159 shortlisted papers and searched for their presence in our database of total number of papers. This extra control brought about a revelation of one new applicable paper that met the previously stated criteria. This paper was inserted to our database for data extraction; henceforth, the number of shortlisted papers expanded to 160. To ensure the legitimacy of the shortlisted papers, the control process was repeated by randomly selecting 32 papers (20%) of the 160 chosen papers which produced 100% relevancy rate with the selection criteria.

The third phase of SLR (i.e. reporting) is discussed in the following sections in which each SLR question mentioned before is recalled and answered.

4. Criteria for optimal features

This section answers the first question:

Question 1: "What are the criteria for optimal features and how feature parameter appropriateness is decided for feature extraction process in SI?"

In order to answer this question the following list of characteristics of optimal features were identified (Hansen & Hasan, 2015; Kinnunen & Li, 2010):

(1) Easily measurable and extractable.
(2) Naturally and repeatedly used words in speech.
(3) Discriminating speakers based on changes (differences) in the common features among various speakers.
(4) Based on features that are difficult to mimic.
(5) Not vary over time and during transmission.
(6) Robust against noise and distortions.
(7) Not altered by background noise or speaker health.
(8) Features that are unique and maximally independent of other features.

The first and second characteristics of the optimal features deal with easing the process of features extraction. Since the features extraction process occurs regularly, it is necessary to make this process simple and easy. The third characteristic is for distinguishing the voices of two speakers with lower intra-speaker and higher inter speaker variations. The fourth to the eighth indicate the robustness of the features extracted. i.e. how features are affected by channel noise, distortions, speaker health and difficulty to mimic or disguise. The last two suggest that extracted features must be independent of each other. This is because when two correlated features are combined then nothing gained rather than degrading the quality of identification of the speaker (Kinnunen, 2003).

Every speaker has unique vocal characteristic features. These features can be classified into two categories namely learned (behaviour based) and physiological features each of which are further classified as shown in Fig. 7.

4.1. Learned (behaviour based) speech features

The viewpoint behind learned speaker features are education, background, parental influence, personality types, place of birth and language. Learned speaker features are further classified into two types of features: high level, and prosodic and spectro-temporal feature.

High level features include phones, idiolect (personal lexicon), semantics, accent, and pronunciation. The following list presents the main characteristics of high level features:

- *Phones:* phones are occasions of phonemes in the genuine expressions (i.e. the physical fragments). Phoneme is the smallest constructional unit that decide significance in a dialect. For example, they are the perceptually distinct units in a specified language that differentiate one word from others like *b* and *v* in English words *berry* and *very*. Phonemes are not the physical fragments themselves, but rather are psychological abstractions or categorization of them.

Table 1
Paper review selection process.

Selection process	Selection criteria	Paper removed	Total
Paper extracted from electronic databases	Search term = "Speaker Identification"	–	535
Screenings based on the year 2011 to 2016	Search term = "Speaker Identification" *and* applying the time limitation	345	190
Removing non-English papers	Search term= "Speaker Identification" *and* applying the time limitation *and* language = "English"	2	188
Filtering based on the title and abstract	Search term = "Speaker Identification" *and* applying the time limitation *and* language = "English" *and* analyzing the title and abstract of the paper	21	167
Paper removed based on other criteria	Search term = "Speaker Identification" *and* applying the time limitation *and* language = "English" *and* analyzing the title and abstract of the paper *and* applying other exclusion criteria	8	159

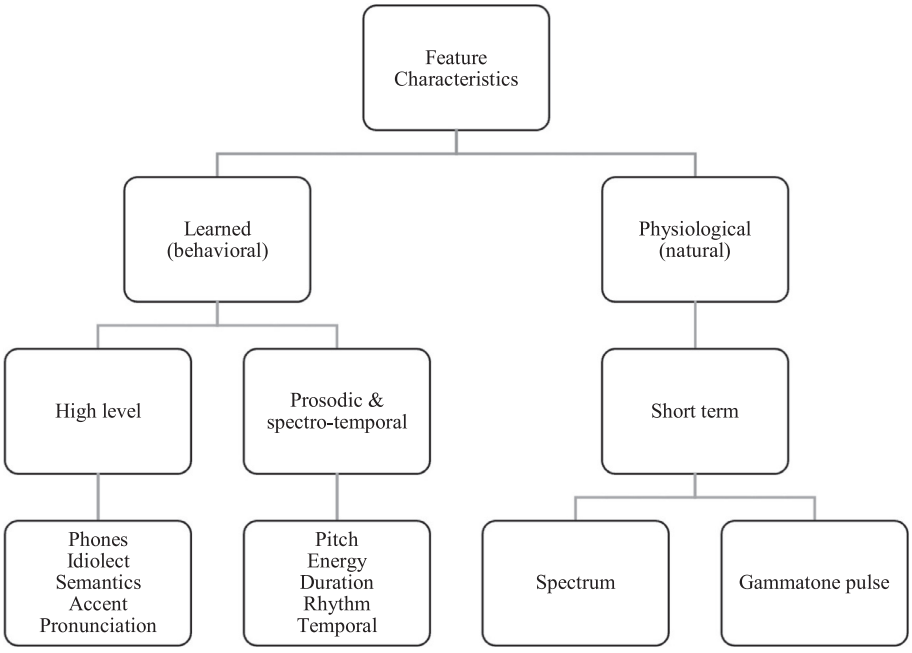

Fig. 7. Speaker feature characteristics classification.

- *Idiolect (personal lexicon):* this is defined as unique speaking manner or habits of a particular speaker. Dialect is a class of a language spoken by a group of people.
- *Semantics:* it is the speaker competence with regards to the meaning of linguistic structures.
- *Accent:* a special force given to a syllable or word in speech by stress or pitch.
- *Pronunciation:* the way in which a word is pronounced.

Prosodic and spectro-temporal features comprise of pitch, energy, duration, rhythm and temporal features. Frequency based features (or spectral features) are extracted by performing time-based signal to the frequency domain conversion. This is usually performed by applying Fourier Transform (FT) techniques such as spectral centroid, spectral flux, fundamental frequency, spectral density, frequency components, etc. (Sarma & Sarma, 2013b). Moreover:

- Pitch is a perceptual attribute of the speaker voice with physical attributes denoted by the fundamental frequency.
- Fundamental frequency (commonly known as F0) is the ratio of vocal fold vibrations during voice phonation. It is pertinent to note that pitch and F0 are both considered when refereeing to vocal fold vibration frequency in the literature although they are different quantities (Kinnunen, 2003).
- Temporal (time domain) features are generally simple to extract and have easy physical interpretation. Examples are signal energy, maximum amplitude, minimum energy, zero crossing rate, etc. (Sarma & Sarma, 2013b).
- Prosodic features are non-segmental aspect of speech found in long utterances, for example prosodic features intonation is a collective term that used to explain the variations in pitch, loudness, rhythm and stress.

Table 2
Chronology of behavioural features extraction approaches.

Feature category	Feature extraction approaches	References
High Level	Seven Shannon entropy wavelet packet and five formants were extracted from vowelM	(Daqrouq & Tutunji, 2015)
	PMFC (Phoneme Mean F-ratio Coefficient)	(Zhao et al., 2012)
	PMFFCC (Phoneme Mean F-ratio Frequency Cepstrum Coefficient)	(Cumani, Plchot, & Karafiát, 2012)
	Polish vowel (stressed or unstressed) in selected contexts described by the four lowest formant frequencies	(Salapa, Trawińska, Roterman, & Tadeusiewicz, 2014)
	Vowel phonemes (LPC with Self organizing map (SOM) for segmentation of vowels)	(Sarma & Sarma, 2013b)
	Maximum-Likelihood Linear Regression (MLLR)	(Saeidi, Hurmalainen, Virtanen, & van Leeuwen, 2012)
	Constrained Maximum-Likelihood Linear Regression (CMLLR) coeffs	
Prosodic and spectro-temporal	Algorithms or methods used for extraction of features Sub-band Auto Correlation Classification (SACC)	(El Khoury, Laurent, Meignier, & Petitrenaud, 2012; Kawakami et al., 2014; Lu, Brush, Priyantha, Karlson, & Liu, 2011; Nagaraja & Jayanna, 2012; Plchot et al., 2013; Prasad, Periyasamy, & Ghosh, 2015; Wu & Tsai, 2011)
	PROSACC	(McLaren, Scheffer, Graciarena, Ferrer, & Lei, 2013)
	Empirical Mode Decomposition (EMD) features extraction method (Calculating the energy of each component reduce computation)	
	Shifted Delta Cepstrum (SDC) and additional temporal information	(Kockmann et al., 2011)
	DPE to present pitch and energy by using twelve DCT coeffs	(Kockmann et al., 2011)

- Short-term spectral features are considered for short duration since the voice signal is continuously changing as the result of articulation. The speech signal is usually broken down into short frames having intervals of 25–30 ms. For such small time the features are considered to be stationary and these small frames are selected for spectral features extraction.
- Mel is known as a unit of perceived fundamental frequency.
- Fast Fourier transform (FFT) is a faster version of DFT that decomposes a signal into frequency components. The global shape within a single frame of the DFT magnitude spectrum is called spectral envelope. The spectral envelope is one of the most informative part of the spectrum of speaker identification that contains the information of resonance properties of vocal tract.

Table 2 provides a chronology of behavioural feature extraction approaches.

4.2. Physiological based speech features

Psychological (i.e. natural) features are influenced by the length and dimension of the vocal track, and the size of vocal folds. Short term spectral features are calculated from short speech frames; they are used for describing the short term spectral envelope correlating to timbre and resonance properties of supralaryngeal vocal tract (which consist of oral, pharyngeal, and nasal cavities). Voice source features are properties of the glottal flow. Prosodic and spectro-temporal span over tens to hundreds of milliseconds and responsible for controlling the intonation, stress, and rhythmic organization of the speech (Pierrehumbert, 1980).

Mel-Frequency Cepstral Coefficients (MFCCs) is a collection of coefficients that are used as features; they are constructed using frequencies of vocal track information. They present acoustic signals in cepstral domain that employ FFT to represent windowed short signals as the real cepstrum of the signal. It is inspired by our natural auditory perception mechanism hence MFCC frequency bands are spaced equally on Mel scale (Shahamiri & Salim, 2014a).

Filter bank based MFCC features extraction method is depicted in Fig. 8. MFCC mainly represents the vocal tract information. Its calculation is based on filter bank method but executed using time frequency analysis. Initially, the time analysis is performed by applying framing operation; this is followed by applying frequency analysis based on progressing the speech frame through filter bank. MFCC needs overlapping frames because time analysis is done in advance. Filter banks are designed in a manner to operate in a similar way to the human auditory frequency perception (Ma & Leijon, 2011). To represent full dynamic feature of MFCCs, dynamic information contained over the time sequence, like the velocity and the acceleration, are usually combined with the MFCCs (Sen & Basu, 2011a).

Short term MFCC based feature extraction methods are shown in Table 3.

The short term spectral features are further classified into two types namely Spectrum and Gammatone pulse features; they are discussed in the next section.

5. Feature extraction approaches

This section deals with the second question:

Question 2: "What are the feature extraction approaches and algorithms used in SI process?"

Table 3
Short term MFCC based feature extraction.

Feature category	Feature extraction approaches	References
Short term and voice source feature	MFCC	(El Khoury et al., 2012; Lu et al., 2011; Plchot et al., 2013; Prasad et al., 2015) (Biagetti, Crippa, Curzi, Orcioni, & Turchetti, 2015; Biagetti, Crippa, Falaschetti, Orcioni, & Turchetti, 2016; Chao, 2012; Esmi, Sussner, Valle, Sakuray, & Barros, 2012; Fan & Hansen, 2011b; Fang & Gowdy, 2013; Fazakis, Karlos, Kotsiantis, & Sgarbas, 2015; Gabrea, 2011; Ghiurcau, Rusu, & Astola, 2011; Gong, Zhao, & Tao, 2014; Hanilcc et al., 2013; Kim, Yang, & Yu, 2013; Li & Huang, 2011; Li, Delbruck, & Liu, 2012; Liu & Guan, 2014; Michalevsky et al., 2011; Mitra, McLaren, Franco, Graciarena, & Scheffer, 2013; Nugraha, Yamamoto, & Nakagawa, 2014; Pal, Bose, Basak, & Mukhopadhyay, 2014; Pathak & Raj, 2013; Prasad, Tan, & Prasad, 2013; Sadjadi & Hansen, 2013; Prasad et al., 2015; Safavi, Hanani, Russell, Jancovic, & Carey, 2012; Sarkar & Umesh, 2011; Sarkar, Umesh, & Bonastre, 2012; Sen & Basu, 2012; Sidorov et al., 2013; Taghia, Ma, & Leijon, 2013; Trabelsi & Ayed, 2014; Wang et al., 2015; Wang, Zhang, & Kai, 2013; Wang, Zhang, Kai, & Kishi, 2012; Xing, Li, & Tan, 2012; Yamada et al., 2013; Yang & Liu, 2014; Yang, Chen, & Wang, 2011; Zao & Coelho, 2011; Zhang, Wang, & Kai, 2014; Zhao & Wang, 2013; Zhao, Wang, & Wang, 2014)

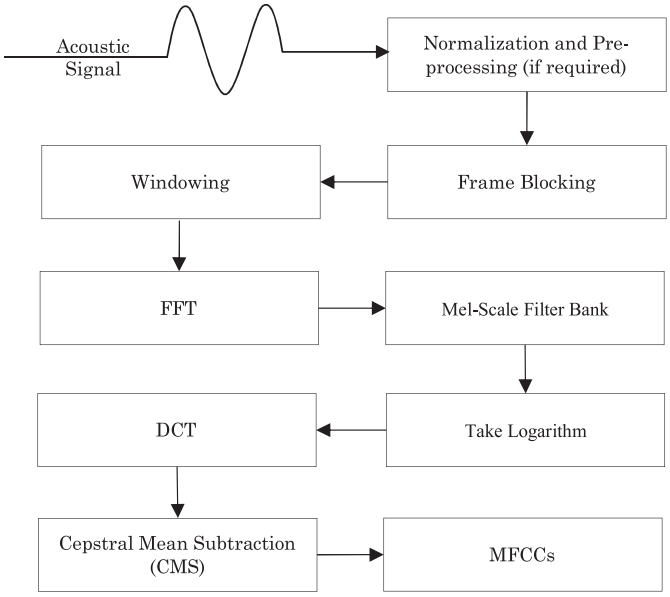

Fig. 8. Filter bank based MFCC feature extraction method.

Speech signals consist of information about human speech production and auditory system. It is important for the extracted speech features to provide adequate discriminative data that appropriately fit with SI back-end modelling. SI approaches mainly tend to extract the vocal track physical attributes that are described as acoustic resonance properties (Zhao, Wang, Hyon, Wei, & Dang, 2012). Various algorithms and methods are reported in the literature to extract the information from speech either by modelling the human voice production system or modelling the peripheral auditory system, each of which is explained in the following.

5.1. Feature extraction approaches to model the human voice production

It is possible to extract features that best represent phonemes by understanding the human voice production system. In this regard, Linear Prediction Coding (LPC) is one of the methods for spectrum feature extraction which provides good interpretation for both time domain and frequency domain. The former means the correlation of all adjacent samples and the last means all pole spectrum referring to the resonance structure. LPC is capable of providing an accurate estimation of speech spectra, formants, and pitch by mimicking the human voice production system. It is commonly popular in SI because of its easy and fast applicability and capability of extracting and storing time-varying formant data. Linear Predictive Cepstral Coefficients (LPCC), derived from LPC, is another popular group of cepstral coefficients in order to model the human voice production system in clean environment that applies a filter to simulate the vocal tract (Malegaonkar & Ariyaeeinia, 2011; Wang, Peng, Wang, Lin, & Kuan, 2011).

Table 4 presents linear predication based approaches. Most of the LPC-based approaches are confined to perceptual modelling. Furthermore, Table 5 provides other methods that employ linear prediction.

5.2. Feature extraction approaches to model the peripheral auditory system

These approaches can either be based on Fourier or auditory transform. The essential standard behind scaling speech signals is that the high-energy areas of speech spectrum have greater part of linguistic data than low energy region. Utilizing this guideline, speech signal-based frequency warping was proposed by considering equivalent range segments based on calculating the ensemble average short-time power spectrum (EAPS) of whole speech corpus logarithm (Sarangi & Saha, 2012).

The general process of speech signal based feature extraction is shown in Fig. 9(a). The pre-processed speech signal is squared with FFT of windowed estimate; the magnitudes of this is called Periodogram that is applied by MFCC for spectrum estimation and calculation of ensemble average of power spectrum followed by calculation of log. Finally, the outcome is divided into equal intervals where central frequencies of each equal area interval is calculated; this is also known as speech-signal based frequency warping function.

In addition, the filter bank based cepstral feature extraction process is shown in Fig. 9(b). The process is similar to 9 (a) except a filter bank is used after the FFT process. The survey on these implementations is presented by Tables 6–8.

Recently there have been some variation and amalgamations with MFCC. For example, Yu, Ma, Li, and Guo (2014) proposed an approach that constructs a MFCC super frame by combining two

Table 4
Linear Prediction based approaches.

Feature extraction approaches	References
Linear Predictive coding-derived Cepstral Coefficient (LPCC)	(Kawakami et al., 2014; Malegaonkar & Ariyaeeinia, 2011; Qi & Wang, 2011; Rossi, Amft, & Tröster, 2012; Wang et al., 2011)
Linear Predictive Coding (LPC)	(Chandra, Nandi, Mishra, & others, 2015; Do, Tashev, & Acero, 2011; Qi & Wang, 2011)
Linear Predictive Residual (LPR)	(Kawakami et al., 2014; Khan, Basu, & Bepari, 2012)
Perceptual Linear Prediction (PLP)	(Bredin, Roy, Le, & Barras, 2014; McLaren et al., 2013; Plchot et al., 2013)
Perceptual Linear Prediction Cepstral Coefficient (PLPCC)	(McLaren et al., 2013; Plchot et al., 2013)
Frequency Domain Linear Prediction (FDLP)	(Godin, Sadjadi, & Hansen, 2013; Plchot et al., 2013)
Wavelet LPC (WLPC)	(Chandra, Nandi, & Mishra, 2015)
Log Area Ratio (LAR) and Perceptual LAR (PLAR)	(Sidorov et al., 2013)
Wavelet Transformation	(Srinivas, Rani, & Madhu, 2014)
Dyadic Wavelet Transform (DWT) DWTLPC (i.e. DWT + conventional LPC) DWTLPCF (i.e. DWT + AFLPC) WPID (Wavelet Packet Energy Index Distribution) GWPNN (Genetic Wavelet Packet ANN) WPLPC (i.e. Wavelet Packet + conventional LPC) WPLPCF MDWTLPC (i.e. Modified DWT + Conventional LPC) EVPLPC (i.e. Eigen vector + conventional LPC + WP) EDWTLPC (i.e. Eigen vector + DWT + LPC)	(Daqrouq & Al Azzawi, 2012)
Peak Difference Auto-correlation of wavelet transform (PDAWT)	(Ghezaiel, Slimane, & Braiek, 2013)
Multi-Resolution Dyadic Wavelet (MRDWT)	(Ghezaiel et al., 2013; Ghezaiel, Slimane, & Braiek, 2012)
Frequency Cepstral (WPMFC) Features)	(Srivastava et al., 2013)
Average Framing Linear Prediction Coding (AFLPC)	(Daqrouq & Al Azzawi, 2012)
Temporal Energy Sub-Band Cepstral Coefficients (TESBCC).	(Sen & Basu, 2011a)
Fourier-Bessel based Cepstral Coefficient (FBCC)	(Prakash & Gangashetty, 2011; Vasudev & K, 2014)
short time Log Frequency Power Coefficients (LFPCs)	(Shahin, 2013)
Complex Cepstrum Temporal Filtering (CCTF)	(Vannicola, Smolenski, Battles, & Ardis, 2011)

Table 5
Other Linear Prediction approaches.

Feature extraction approaches	References
LP and EMD	(Dutta et al., 2015)
Root Mean Square (RMS) (on various domains)	(Fernando, Ramey, & Salichs, 2014)
Spectral and Residual Features	(Sahidullah & Saha, 2011)
MFCC at leaf node only + Pitch + Five feature extracted from LP residual signal that *"are width of the positive pulse, skewness of the positive pulse, skewness of the negative pulse, PAR of the positive pulse within one cycle and PAR of the negative pulse within one cycle"*	(Hu, Wu, & Nucci, 2013)
ACWPFL (i.e. Adaptive Component Weighted cepstrum Post-Filter) CEP (i.e. Linear Predictive Cepstrum) PFMRCEP (Pole Filtered Mean Removed Cepstrum) PFMRACW (Pole Filtered Mean Removed ACW) cepstrum + MRPFL (Mean Removed PFL cepstrum) Mean Removed ACW cepstrum (MRACW)	(Ramachandran, Polikar, Dahm, & Shetty, 2012)
SDC (Shifted delta cepstral) + PMVDR (Perceptual Minimum Variance Distortion Response)	(Liu, Lei, & Hansen, 2012)
LP Coefficient based	(Li et al., 2012; Raval, Ramachandran, Shetty, & Smolenski, 2012; Sarangi & Saha, 2012)
Using EMD (i.e. Empirical Mode Decomposition) to extract speaker's physiologically motivated features (i.e. glottal source information). Examples are RPCC (Residual Phase Cepstrum Coeffs), TPCC (i.e. Teager Phase), and GLFCC (i.e. Glottal Flow)	(Sarma & Sarma, 2013a; Wang & Johnson, 2014)

Table 6
MFCC variation and fusion based feature extraction chronology.

Feature Category	Feature extraction approaches	References
Variation of MFCC	Super MFCC	(Yu et al., 2014)
	Log based MFCC	(Jawarkar et al., 2015)
	Cubic root based MFCC	
	MFCC FB 24	(Davis & Mermelstein, 1980).
	MFCC FB 26 HTK	(Young et al., 2006)
	MFCC FB40	(Bouziane, Kharroubi, & Zarghili, 2014)
	SWCE (i.e. Sine-Weighted Cepstrum Estimator) taper MFCCs	(Nagaraja & Jayanna, 2012)
	Multitaper MFCC	(Nagaraja & Jayanna, 2013)
	20th-order regularized LP-MFCCs (RLP)	(Godin et al., 2013)
	MFCC of LP Residual	(Kawakami et al., 2014)
	Multi frame rate MFCC	
	Different weights to ignore low energy frame and emphasizing on high energy frame of MFCC	(Ayoub et al., 2014)
	Inverted Mel Frequency Cepstral Coefficients (IMFCC)	(Sen & Basu, 2011b)
	RASTA	(Kockmann et al., 2011)
	RASTA-D	
	RASTADD	
	RASTA-DDD	
	SDC (Shifted delta cepstral)	
	DPEC	
	Mel frequency log filter bank power spectrum	(Srinivasan, Ming, & Crookes, 2012)
	Reversed Mel-frequency cepstral Coefficients (RMFCC)	(Do et al., 2011)
MFCC fusion	MFCC + time delay of arrival(TDOA)	(Gong et al., 2014; Kawakami et al., 2014; Kawakami et al., 2013; Sadjadi & Hansen, 2013; Schmidt et al., 2014; Yang & Liu, 2014)
	De-reverb MFCC + TDOA	(Yang & Liu, 2014)
	De Reverb MFCC	
	MFCC with Local sensitive hashing (LSH)	(Tomar & Rose, 2013)
	MFCC + Bark	(Zhang, Bai, & Liang, 2006)
	FFT-MFCC and LP-MFCC	(Godin et al., 2013)
	Multitaper MFCCs + LPR + LPRP	(Nagaraja & Jayanna, 2013)
	MFCC and wavelet transform	(Verma, 2011)
	MFCC and phase information	(Nakagawa, Wang, & Ohtsuka, 2012)
	MFCC and MHEC	(Sadjadi & Hansen, 2013)
	MFCC +LPCC (vocal tract feature)	(Malegaonkar & Ariyaeeinia, 2011)
	LPCC+ LPC Residual	(Islam & Rahman, 2009)
	MFCC+RPCC	(Wang & Johnson, 2014)
	MFCC+GLFCC	
	MFCC+TPCC	
	MFCC+RPCC+GLFCC+TPCC	
	MFCC + Histogram Transform	(Ma, Yu, Tan, & Guo, 2016)
Other filter bark based approaches	Extraction Algorithm Bark	(Gong et al., 2014; Mitra et al., 2013)
	Bark Spectral Flatness Measure (BSFM)	(Gong et al., 2014)
	Bark Spectral center (BSC)	
	MDMC (Medium Duration Modulation Cepstrum) MMeDuSA (i.e. Modulation features of Medium Duration sub-band Speech Amplitudes)	(Mitra et al., 2013)

neighboring frames of a current frame. By using MFCC super frame the probability density function was calculated that diminishes the discontinuity problem of the common multivariate histograms (Lyubimov, Nastasenko, Kotov, & Doroshin, 2014). Table 6 presents the details of our literature survey on various MFCC based approaches that are developed by variations, fusion of MFCC approaches, and other MFCC based feature extraction approaches.

The most difficult part in feature extraction is to extract dissimilar features known as bottleneck features. MFCC has been successful in extracting bottleneck features for distant-talking speaker identification (Yamada, Wang, & Kai, 2013). Furthermore, Multilayer Perception (MLP) Artificial Neural Networks (ANNs) were employed for extracting bottleneck features successfully too (Matejka et al., 2014). Bottleneck features provided by MLPs are applicable for transforming nonlinear features and reducing their dimensions. Particularly, an MLP was trained using the feedforward and backpropagation algorithm by selecting initial weights and biases randomly. The training process then followed by reducing the dimensions of several frames of cepstral coefficients (Qi, Wang, Xu, & Tejedor Noguerales, 2013). It was shown that using and integration of bottleneck features together with the coefficients provided better results than conventional pure MFCCs.

In addition to MLP based traditional ANNs, Deep Neural Networks (DNNs) with layer-wise training attained state of the art results for various machine learning problems. There are also several implementations of deep learning for both enrolment and match-

Table 7
Short term and voice source feature: Spectrum.

Feature extraction approaches	References
WT TEOCC (Teager Energy Operator based coeffs) MDMC RT + DCT	(Ajmera, Jadhav, & Holambe, 2011; Daqrouq, 2011; Deshpande & Holambe, 2011b; EhKan, Allen, & Quigley, 2011; McLaren et al., 2013; Shih, Lin, Wang, & Lin, 2011)
FF-ratio Frequency Cepstrum Coefficient (FFCC), 10-dimensional cepstral vector, frame level log-spectral features	(Ding & Yen, 2015; Hyon, Wang, Zhao, Wei, & Dang, 2012; Zhao, Wang, & Wang, 2015)
MFCC combined with the following masking: forward + temporal + lateral inhibition	(Wang, Tang, & Zheng, 2012)
BF-DNN / BF-MLP: bottleneck features extracted from MFCC 1-frame of 25-dimensional MFCC features that were inputted to DNNs with pretraining Denoising Autoencoder (DAE) based cepstral-domain reverberation	(Yamada et al., 2013; Zhang et al., 2015)
DCT coeffs histogram	(Al-Rawahy, Hossen, & Heute, 2012)
For compensating whispered, soft, neutral, loud and shouted voices features interconversion	(Hanilcc et al., 2013)
MFSC (Mel Frequency Spectral Coefficients)	(Ouamour & Sayoud, 2013)
LFCC (linear frequency cepstral coefficients)	(Jourani, Daoudi, Andre, & Aboutajdine, 2013; Sahidullah et al., 2011; Sidorov et al., 2013)

Table 8
Other Short term feature based approaches.

Feature extraction approaches	References
RASTA-PLP	(Li & Gao, 2016; Li & Huang, 2011; Trabelsi & Ayed, 2014)
Gammatone Feature	(Zhang, Zhang, & Gao, 2014; Zhao et al., 2014; Zhao, Shao, & Wang, 2011, 2012)
Gammatone frequency cepstral coefficients (GFCC)	(Jawarkar et al., 2015; Li & Huang, 2011; Zhang et al., 2014; Zhao et al., 2011, 2012; Zhao et al., 2014)
Hilbert envelope of Gammatone filterbank	(Sadjadi & Hansen, 2011)
Mean Hilbert envelope coefficient (MHEC)	(Godin et al., 2013; McLaren et al., 2013; Sadjadi & Hansen, 2013; S.O. 2015)
Power-normalized cepstral coefficient (PNCC)	(McLaren et al., 2013; Mitra et al., 2013; Sadjadi & Hansen, 2015)
Auditory-based, time frequency Transform (AT)	(Li & Huang, 2011; Plchot et al., 2013)
Fisher Vector (FV)	(Jiang, Frigui, & Calhoun, 2015)

ing phases for SI. In some approaches, deep learning was used for feature extraction whereas in other approaches only for matching as a classifier (Dutta et al., 2015; Justin et al., 2015; LeCun, Kavukcuoglu, & Farabet, 2010; Pobar & Ipsic, 2014; Xie, Xu, & Chen, 2012).

An example of using DNNs to perform classifications is Lukic, Vogt, Dürr, and Stadelmann (2016) where an SI system based on CNNs was proposed and verified using connected speech samples provided by TIMIT. The network employed 32 and 64 filters for its convolutional layers each followed by another layer to perform max-pooling. The final layers were two dense layers.

DNNs were also employed to deal with the bottleneck features. In particular, Dutta and others applied a five-layer DNNs with 3-frames of 25 dimensional MFCC features as the input to the DNNs (Dutta et al., 2015). There were 25 hidden units in the bottleneck layer and 500 hidden units in non-bottleneck hidden layer. The MFCC normalization was performed considering the training data mean. The DNN training process was performed by applying stochastic mini-batch gradient descent that had a minibatch size of 100 samples. Initially a pre-training process of fifty cycles with learning rate 0.1 was conducted. Next, another training was conducted following up the pre-training process in which 1000 training cycles were performed with the same learning rate. Three kinds of methods were compared in this study:

- Method 1 in which a GMM was trained using MFCC and applied as a baseline system (denoted as MFCC).
- Method 2 that used the bottleneck features extracted from the MLP without pre-training (denoted as BFMLP).
- Method 3 that employed the bottleneck features extracted from the DNNs with pre-training (denoted as BF-DNNs).

DNNs with pre-training delivered a better performance than the conventional MLP without pre-training when it was used in ASR systems with a large vocabulary size (Yamada et al., 2013). This factor-analysis-based framework incorporated a vector extractor module with a Bayesians probabilistic linear discriminant anal-

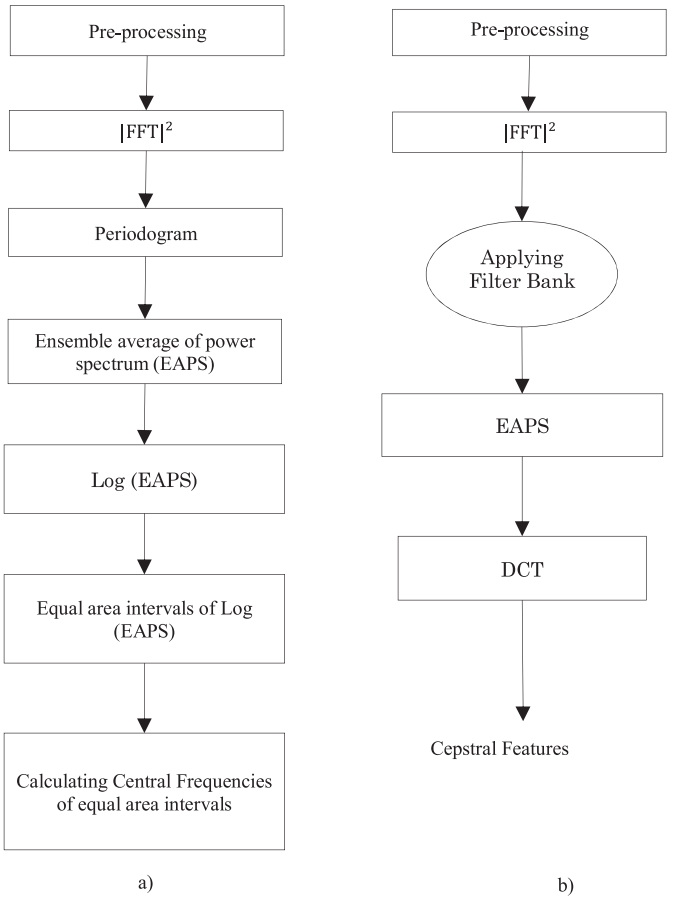

Fig. 9. (a) Speech signal based scaling process, (b) Feature extraction process using filter bank (Sarangi & Saha, 2012).

ysis (PLDA). Another use of DNNs in extracting bottleneck features was also studied by Matějka et al. (2016).

In addition to MFCC, another FT-based popular feature is Relative spectra Filtering-Perceptual Linear Predictive Coefficients (RASTA-PLP) (Kockmann, Burget, & Cernocký, 2011). For the cochlear model, Gammatone filter banks were used to perform auditory transform (Li & Huang, 2011). Note that the method of extracting the features mainly depends on the associated feature domain; they are the high level feature domain, prosodic and spatiotemporal feature domain, short term spectral and voice source feature domain.

5.3. Other feature extraction approaches

Our research is further extended to understand other methods that may not directly come under the previous categories. They are presented in Tables 9 and 10.

6. Feature extraction approaches for last six years

In this section, we present the results of our literature survey on various approaches for features extraction to answer the last question:

Question 3: "What is the most popular and successful feature extraction approach in the last six years?"

Fig. 10 illustrates the contribution distribution of different SI features extraction approaches in the past six years. As can be seen, MFCC-based feature extraction approaches were employed in SI more than any other methods mentioned in the literature. In particular, using a form of MFCC was reported in 97% of the publications during the last six years. We divided the literature review on the MFCC based approaches into six different categories: pure MFCC, short-term based, MFCC variations, MFCC fusion, linear prediction, and other MFCC approaches.

Likewise, the SI systems that applied MFCC-based approaches reported to obtain better results in compare to the rest. The literature on the MFCC-based approaches was already provided in the last section.

Nevertheless, despite the popularity of the MFCC-based approaches the performance of SI systems using them can drop, specifically when speech is infected with noise, due to the complex nature of real-time speech data. Thus, cleansing algorithms were reported to be useful to improve this situation. Approaches and algorithms in Table 11 were reported for enhancing the features extracted using MFCC.

7. Discussion

According to the literature and popularity of MFCC-based approaches, we categorized the features extraction approaches into MFCC based and non-MFCC based. The MFCC based feature extraction approaches were further classified into the following approaches:

- Pure MFCC
- MFCC Variations
- MFCC Fusion
- Short Feature based
- Linear Prediction
- Other MFCC

As discussed, the MFCC-based feature extraction approaches were identified as the most popular, successful and widely used approach for SI feature extraction with 97% of overall implementations in which 31% being pure MFCC. There was a continues raise especially from 2013 onwards in publications using approaches that incorporate MFCC by enhancing the MFCC features or amalgamating MFCC approaches with other methods. From the literature review it can be observed that the majority of the machine learning based implementations of speaker identification used a form of the MFCC feature extraction process too. However, number of publications have not increased in respect to studies employing pure MFCC approaches (i.e. without amalgamations). This shows that the research trend is towards combination of various approaches. Further analysis also proves the importance of cleansing activities on the extracted features using MFCC in order to improve the speaker identification performance. MFCC coefficient values fluctuate depending on the type of filtering methods (Luengo et al., 2008). Our study shows that most of the state-of-the-art SI systems use MFCC as feature extractor, and then feed these features to GMM-based approaches for creating speaker models for identification. When speaker identification approaches are modelled using statistical methods with features extracted using MFCC, filtering approaches for SI are not affected. One such filtering approach is RelActive SpecTrAl (RASTA) (Kockmann et al., 2011) where MFCC individual cepstral coefficients are filtered.

Furthermore, generating a new feature set by combining features using diarization data driven model, like HMM, produced efficient results compared with traditional MFCC based feature extraction. However, these approaches are also categorized as MFCC since the new features set is a subset of MFCC features with less number of dimensions (Wang et al., 2011).

A diffusion-map based approach with MFCC features for SI is also proven to be efficient (Michalevsky, Talmon, & Cohen, 2011). Approaches like Cubic-root based MFCC and MFCC FB-40 were successfully created by changing MFCC feature extraction methods.

Table 9
Other characters considered as features for speaker identification.

Feature extraction approaches	References
The authors proposed an approach based on characteristics of facial muscles that are involved in lip movements that adopted the muscles intrinsic properties extracted from dynamic lip simulation. Examples of the muscles properties are mass, elasticity, and viscosity. Another example is using lip texture for speaker identification in which distributed nature of lip texture representation was used to discriminate the speakers.	(Asadpour, Homayounpour, & Towhidkhah, 2011; Lai, Wang, Shi, & Liew, 2014; Meng, Hu, Zhang, & Wang, 2011)
Using thirty geometrical features, an SI system was proposed based on lip biometric features by applying MRMR (Minimum Redundancy Maximum Relevance) that reduced the number of visual features.	(Singh, Laxmi, & Gaur, 2012)
Local Spatiotemporal Directional Feature (LSDF)	(Zhao & Pietikäinen, 2013)
Encoding the shape of the mount was employed for SI using MBH (i.e. (Motion Boundary Histograms).	(Rekik, Ben-Hamadou, & Mahdi, 2015)
Joint Factor Analysis (JFA)	(Deshpande & Holambe, 2011a; Yang, et al., 2011)
Line Spectral Frequencies (LSFs) Differential LSF (DLSF)	(Ahmed et al., 2015; Almaadeed, Aggoun, & Amira, 2012; Ma & Leijon, 2011)

Table 10
Other feature extraction approaches.

Feature extraction approaches	References
VTS (Vector Taylor Series) and CMLLR (Constrained Maximum Likelihood Linear Regression) were used to generate neutral features. VTTP (Vocal Tract Tube Profile) Articulation Style (ARTS) Vocal Tract Length (VTL) Log Frame Energy (LOGE) Zero Crossing Rate (ZCR) Spectral Entropy (SE)	(Ethridge & Ramachandran, 2015; Fan & Hansen, 2011a; Mizobe et al., 2012; Prasad et al., 2015)
Binarised voice biometric template were provided by creating speech vectors comprised of the twelve MFCCs + log energy + their first and second derivations. Post-processing of the speech vectors was also conducted by calculating cepstral mean subtraction and feature warping.	(Sahidullah et al., 2011)
Supervector	(Trabelsi & Ayed, 2014)
GMM Supervector	(Kundu, Das, & Bandyopadhyay, 2012; Xing, et al., 2012)
Textual features	(Kundu et al., 2012)
Secondary features	(Saeidi, et al., 2012)
Locality Sensitive Hasing (LSH)	(Godin et al., 2013; T. Liu & Guan, 2014; Plchot et al., 2013; Saeidi et al., 2012; Schmidt et al., 2014)
Dominant Speaker Identification	(Vandyke, Wagner, & Goecke, 2013; Volfin & Cohen, 2013)
Glottal Closure Instant (GCI)	(Vandyke et al., 2013)
Kernel Partial Least Squares (KPLS)	(Bakry & Elgammal, 2013)

Another successful approach with variation of MFCC is multi frame rate MFCC where new variables were introduced to ignore low energy frequency frames in order to emphasize on the high energy frames. Inverted MFCC is also another successful variant of MFCC (Kim, Yang, & Yu, 2013).

Another implementation was a self-organizing mixture models which replaced the EM algorithm of MFCC with self-organizing maps that provided better results than typical GMM-based models of SI (Ayoub, Jamal, & Arsalane, 2014). Similarly, the MFCC delta phase features extraction approach (MFDC) employed a similar process as MFCC (with 13 coefficients) but outperformed typi-

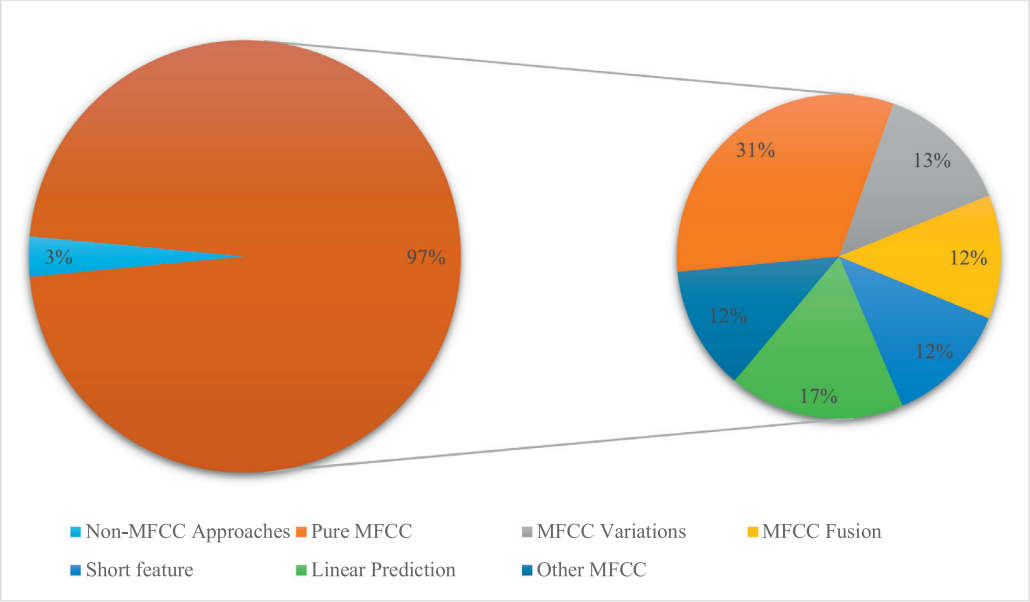

Fig. 10. Survey results for various feature extraction approaches.

Table 11
MFCC enhancement algorithms.

Feature extraction approaches	References
PCA (Principal Component Analysis) Linear Discriminant Analysis or LDA KPCA (Kernel Principal Component Analysis) GKPCA (Greedy Kernel Principal Component Analysis) and KMDA (Kernel Multimodal Component Analysis) ICA (Independent Component Analysis) Principal Component Transformation	(Yang, Kim, So, Kim, & Yu, 2012)
Reducing the features dimension by using DKLT (i.e. Discrete Karhunen-Love Transform)	(Biagetti et al., 2015)
Joint density GMM mapping for compensating the MFCC	(Fan & Hansen, 2013)
Pseudo-whisper features created using MFCC domain by convolutional transformation (ConvTran) CMLLR (Constrained Maximum Likelihood Linear Regression) FA (Factor Analysis) from neutral to whispered speech	
Using BSW technique to pre-process MFCC	(Lyubimov et al., 2014)
MFCC feature enhanced by NMF Linear Constraint (LC)-NMF	
Minimum Mean Squared Error (MMSE)	(Godin et al., 2013)

cal MFCC based approaches due to implementation of longer frame lengths with almost 9.32% more error rate (Schmidt, Sharifi, & Moreno, 2014). RASTA based approaches, such as RASTA-DD and RASTA-DDD, were also reported to outperform pure MFCC feature extraction results (Deshpande & Holambe, 2011b; Kawakami, Wang, & Nakagawa, 2013).

Among the MFCC fusion approaches the most prominent one is Vector Quantization (VQ). The VQ and non-negative matrix factorization (NMF) approach for feature extraction was developed using MFCC-based distance measure instead of conventional approaches which provided better and clearer speech models that resulted in higher recognition rates. Linear prediction based MFCC fusion approaches especially Perceptual Linear Prediction Cepstral Coefficient (PLPCC) and Frequency Domain Linear Prediction (FDLP) methods using Cochlear Filter Cepstral Coefficients (CFCCs) were also outperformed regular MFCC feature based approaches (Plchot et al., 2013).

Another prominent MFCC fusion approach is CFCC which is an auditory based feature extraction method that can address acoustic mismatch that exists between training and testing data. CFCC is also outperformed perceptual linear predictive (PLP). The features extracted with CFCC are also proven better than the RASTAPLP approach. Another approach called Vector Taylor Series (VTS) can be implemented for generating features to address the environmental additive noise which is also the key aspect of CFCC (Fan & Hansen, 2011a). Further extension to VTS like Vocal Tract Tube Profile (VTTP), articulation style (ARTS), Vocal Tract Length (VTL) have also provided promising results (Ethridge & Ramachandran, 2015;

Fan & Hansen, 2011a; Mizobe, Kurogi, Tsukazaki, & Nishida, 2012). Other successful approaches include Gammatone Frequency Cepstral Coefficients (GFCC) (Zhao et al., 2012) and Power-Normalized Cepstral Coefficient (PNCC) (Sadjadi & Hansen, 2015).

Linear prediction based approaches such as Adaptive Component Weighted (ACW), Linear Predictive Cepstrum (LPC), Cepstrum Post-Filter (PFL), Mean Removed ACW cepstrum (MRACW), Pole Filtered Mean Removed Cepstrum (PFMRCEP), Mean Removed PFL cepstrum (MRPFL), and Pole Filtered Mean Removed ACW cepstrum (PFMRACW) were also proven efficient. DWT methods amalgamated with LPC based approaches resulted in a variety of successful implementations like DWTLPC, DWTLPCF, WPLPC, MDWTLPC etc.

Although the implementation strategy is based on the SI applications, the importance of machine learning algorithms to perform the matching phase should not be neglected. Nevertheless, the majority of features extraction approaches employ a form of MFCC irrespective of the algorithm employed for the machining phase. Observing the recent SI literature, it can be concluded that using i-vector based approaches with deep learning is receiving attention since this technique reported to produce significant results for SI.

Neural networks were widely used as the identification algorithm. For example, it was reported that using Wavelet Transform (WT) and neural networks work efficiently for text-independent SI in which speech features represented by Teager Energy Operator based Cepstral Coefficients (TEOCC) and MFCC (Daqrouq, 2011). Recent neural network advances especially in deep learning algorithms have improved the SI recognition rates when speech features were extracted using MFCC.

We have seen a decrease of using the traditional, basic, features extraction approaches without extensions or amalgamations in recent years. In the contrary we have seen an increase in methods that design noise robust SI systems by reducing the impact of noise using data cleansing approaches (Yamada et al., 2013). This is noticeable based on the number of implementations and citations that highlight the significance of using MFCC based feature extraction approaches for data cleansing.

There is a clear indication on requirements and necessity of implementing language-independent SI systems as well. Such systems are expected to identify speakers with training and testing samples being recorded in different languages. Furthermore, recent research on transfer learning has a good prospect for SI systems where the knowledge on the feature characteristics can be transferred to a different type of implementation.

8. Conclusion

This paper presents a systematic review on various features extraction methods and algorithms for speaker identification. We used scholarly recommendation approaches for extracted publications from various sources. We presented the general SI process followed by a detailed survey on various features extraction processes. The importance of identifying significant features and their influence on speaker identification accuracy was also discussed. Around 190 publications between the period of 2011 and 2016 using the Kitchenham systematic review methodology were considered in this study. Initially we constituted a database by extracting 535 papers followed by a four-level filtering process and applying criteria like period, language etc. The in-depth scrutiny of these papers resulted in 160 publications which we studied for this literature review. Reviewing these papers, it can be seen that the majority of features extraction approaches employ a form of MFCC irrespective of the algorithm employed for speaker classification phase.

There is no comprehensive approach recommended in the literature to construct strict recommendations. Hence, we employed recommendations based on approaches incorporated in various scholarly articles. This literature review can serve as a resource for features extraction with respect to speaker identification.

Based on the insights provided by this study, the following future research directions are recommended:

1. It is noteworthy to observe that there is no generalized universal features exaction approach and our research stresses its necessity. However, many research problems intend to maintain a trade-off between the speaker identification accuracy and robustness to noise, which still needs more attention. We recommend future studies investigate the development of a robust universal framework for speaker identification that addresses the important problems of SI. We suggest this universal framework to be:
   - Easily adaptable
   - Capable of modelling multiple languages
   - Allow portable implementation
   - Incorporate the capability to deal with all channel data and noisy data
2. Deep learning technologies have an unprecedented domination in various pattern recognition systems. Future studies should further investigate their applications in SI as both the feature extractor and classifier, and seek how they can contribute towards the universal SI framework.
3. Recent advances in active learning theories, such as Multi-View Enhanced Multi-Learner model (Shahamiri & Salim, 2014b; Shahamiri, Kadir, Ibrahim, & Hashim, 2012), have shown promising results in handling complex pattern recognition problems and improving their efficiencies. Future studies need to investigate whether they can be used to create deep learning based, offline speaker identification models.
4. There has been very limited work on unsupervised speaker identification approaches which involve creating speaker models based on features extracted from unlabeled data. In this respect, approaches like principle component analysis, k-means, factor analysis have not been widely explored in order to propose a technical framework for unsupervised SI.
5. Likewise, extracting speaker distinguishable features from incomplete, tampered or damaged data has not been studied properly despite their wide applications in forensic sciences and data recovery. It is necessary to systematically study how existing SI approaches perform when they are given such tampered data, and investigate methods to improve their performance.

Acknowledgement

This work was supported in part by the Marsden Fund (2013–2017), New Zealand and the NSC Science for Innovation Seed Project (2017).

References

Ahmed, M. Y., Kenkeremath, S., & Stankovic, J. (2015). Socialsense: A collaborative mobile platform for speaker and mood identification. In *Wireless sensor networks: 8965* (pp. 68–83).

Ajmera, P. K., Jadhav, D. V., & Holambe, R. S. (2011). Text-independent speaker identification using Radon and discrete cosine transforms based features from speech spectrogram. *Pattern Recognition, 44*, 2749–2759 *6*.

Al-Rawahy, S., Hossen, A., & Heute, U. (2012). Text-independent speaker identification system based on the histogram of DCT-cepstrum coefficients. *International Journal of Knowledge-based and Intelligent Engineering Systems, 16*, 141–161 *6*.

Almaadeed, N., Aggoun, A., & Amira, A. (2012). Audio-Visual feature fusion for speaker identification. In *Proceedings of the 19th international conference on Neural Information Processing* (pp. 56–67). Doha, Qata: Springer.

Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., & Vinyals, O. (2012). Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing, 20*, 356–370 *6*.

Asadpour, V., Homayounpour, M. M., & Towhidkhah, F. (2011). Audio-visual speaker identification using dynamic facial movements and utterance phonetic content. *Applied Soft Computing, 11*, 2083–2093 *6*.

Atal, B. S. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *The Journal of the Acoustical Society of America, 55*, 1304–1312 *6*.

Ayoub, B., Jamal, K., & Arsalane, Z. (2014). Self-organizing mixture models for text-independent speaker identification. In *2014 Third IEEE international colloquium in information science and technology (CIST)* (pp. 345–350). Tetouan, Morocco: IEEE.

Bakry, A., & Elgammal, A. (2013). Mkpls: Manifold kernel partial least squares for lipreading and speaker identification. In *2013 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 684–691). Portland, OR, USA: IEEE.

Beigi, H. (2011). Speaker Modeling. In *Fundamentals of speaker recognition* (pp. 525–541). Boston, MA: Springer US.

Biagetti, G., Crippa, P., Curzi, A., Orcioni, S., & Turchetti, C. (2015). Speaker identification with short sequences of speech frames. In *ICPRAM 2015 proceedings of the international conference on pattern recognition applications and methods: 2* (pp. 178–185).

Biagetti, G., Crippa, P., Falaschetti, L., Orcioni, S., & Turchetti, C. (2016). Robust speaker identification in a meeting with short audio segments. In I. Czarnowski, A. M. Caballero, R. J. Howlett, & L. C. Jain (Eds.), *Intelligent decision technologies 2016: proceedings of the 8th kes international conference on intelligent decision technologies (KES-IDT 2016) – Part II* (pp. 465–477). Cham: Springer International Publishing.

Boujelbene, S. Z., Mezghanni, D. B. A., & Ellouze, N. (2009). Robust text independent speaker identification using hybrid GMM-SVM System. *International Journal of Digital Content Technology and its Applications, 3*, 103–110.

Bouziane, A., Kharroubi, J., & Zarghili, A. (2014). Self-organizing mixture models for text-independent speaker identification. In (pp. 345–350).

Bredin, H., Roy, A., Le, V.-B., & Barras, C. (2014). Person instance graphs for mono-cross and multi-modal person recognition in multimedia data: Application to speaker identification in TV broadcast. *International journal of multimedia information retrieval, 3*, 161–175.

Busso, C., Hernanz, S., Chu, C.-W., Kwon, S.-i., Lee, S., & Georgiou, P. G. (2005). Smart room: Participant and speaker localization and identification. *IEEE International Conference on Acoustics, Speech, and Signal Processing*: 2. IEEE.

Campbell, J. P. (1997). Speaker recognition: A tutorial. *Proceedings of the IEEE, 85*, 1437–1462 *6*.

Champiri, Z. D., Shahamiri, S. R., & Salim, S. S. B. (2015). A systematic review of scholar context-aware recommender systems. *Expert Systems with Applications, 42*, 1743–1758 *6*.

Chandra, M., Nandi, P., & Mishra, S. (2015). Spectral-subtraction based features for speaker identification. In *Proceedings of the 3rd international conference on frontiers of intelligent computing: Theory and applications (FICTA): 6* (pp. 529–536).

Chao, Y.-H. (2012). Speaker identification using pairwise log-likelihood ratio measures. In *2012 9th international conference on fuzzy systems and knowledge discovery (FSKD)* (pp. 1248–1251). Sichuan, China: IEEE.

Chavan, M., & Chougule, S. (2012). Speaker features and recognition techniques: A review. *International Journal of Computational Engineering Research, 2*, 720–728.

Cumani, S., Plchot, O., & Karafiát, M. (2012). Independent component analysis and MLLR transforms for speaker identification. In (pp. 4365–4368).

Daoudi, K., Jourani, R., Andre, O. R. e. g., & Aboutajdine, D. (2011). In *Speaker identification using discriminative learning of large margin GMM: 6* (pp. 300–307). Springer.

Daqrouq, K. (2011). Wavelet entropy and neural network for text-independent speaker identification. *Engineering Applications of Artificial Intelligence, 24*, 796–802.

Daqrouq, K., & Al Azzawi, K. Y. (2012). Average framing linear prediction coding with wavelet transform for text-independent speaker identification system. *Computers & Electrical Engineering, 38*, 1467–1479 *6*.

Daqrouq, K., & Tutunji, T. A. (2015). Speaker identification using vowels features through a combined method of formants, wavelets, and neural network classifiers. *Appl. Soft Comput., 27*, 231–239.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on, 28*, 357–366.

Deshpande, M. S., & Holambe, R. S. (2011a). Robust speaker identification in babble noise. In *Proceedings of the international conference & workshop on emerging trends in technology* (pp. 635–640). Mumbai, Maharashtra, India: ACM.

Deshpande, M. S., & Holambe, R. S. (2011b). Robust speaker identification in the presence of car noise. *International Journal of Biometrics, 3*, 189–205.

Ding, Jr., & Yen, C.-T. (2015). Enhancing GMM speaker identification by incorporating SVM speaker verification for intelligent web-based speech applications. *Multimedia Tools and Applications, 74*, 5131–5140.

Dişken, G., Tüfekçi, Z., Saribulut, L., & Çevik, U. (2017). A review on feature extraction for speaker recognition under degraded conditions. *IETE Technical Review, 34*, 321–332.

Do, H., Tashev, I., & Acero, A. (2011). A new speaker identification algorithm for gaming scenarios. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP): 6* (pp. 5436–5439).

Dutta, M., Patgiri, C., Sarma, M., & Sarma, K. K. (2015). Closed-set text-independent speaker identification system using multiple ANN classifiers. In *Proceedings of the 3rd international conference on frontiers of intelligent computing: Theory and applications (FICTA) 2014* (pp. 377–385). Springer.

EhKan, P., Allen, T., & Quigley, S. F. (2011). FPGA implementation for GMM-based speaker identification. *International Journal of Reconfigurable Computing, 2011*, 3 *6*.

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition, 44*, 572–587.

El Khoury, E., Laurent, A., Meignier, S., & Petitrenaud, S. (2012). Combining transcription-based and acoustic-based speaker identifications for broadcast news. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4377–4380). Kyoto, Japan: IEEE.

Esmi, E. a. o., Sussner, P., Valle, M. E., Sakuray, F., & Barros, L. c. (2012). Fuzzy associative memories based on subsethood and similarity measures with applications to speaker identification. In (Vol. 6, pp. 479-490).

Ethridge, J., & Ramachandran, R. P. (2015). Rank-based frame classification for usable speech detection in speaker identification systems. In *2015 IEEE international conference on digital signal processing (DSP)* (pp. 292–296). Singapore, Singapore: IEEE.

Ezequiel, L.-R. (2014). A Histogram Transform for ProbabilityDensity Function Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 36*, 644–656 *6*.

Fan, X., & Hansen, J. H. L. (2011a). Speaker identification for whispered speech using a training feature transformation from neutral to whisper. In *INTERSPEECH* (pp. 2425–2428).

Fan, X., & Hansen, J. H. L. (2011b). Speaker identification within whispered speech audio streams. *IEEE Transactions on Audio, Speech, and Language Processing, 19*(6), 1408–1421.

Fan, X., & Hansen, J. H. L. (2013). Acoustic analysis and feature transformation from neutral to whisper for speaker identification within whispered speech audio streams. *Speech communication, 55*, 119–134.

Fang, E., & Gowdy, J. N. (2013). New algorithms for improved speaker identification. *International Journal of Biometrics, 5*(6), 360–369.

Farhood, Z., & Abdulghafour, M. (2010). Investigation on model selection criteria for speaker identification. In *2010 International symposium in information technology (ITSim): 2–6* (pp. 537–541). Kuala Lumpur, Malaysia: IEEE.

Farrell, K. R., Mammone, R. J., & Assaleh, K. T. (1994). Speaker recognition using neural networks and conventional classifiers. *IEEE Transactions on speech and audio processing, 2*, 194–205.

Fazakis, N., Karlos, S., Kotsiantis, S., & Sgarbas, K. (2015). Speaker identification using semi-supervised learning. In *17th International conference of speech and computer* (pp. 389–396). Athens, Greece: Springer.

Fernando, A. M., Ramey, A., & Salichs, M. A. (2014). Speaker identification using three signal voice domains during human-robot interaction. In *HRI '14 Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction* (pp. 114–115). Bielefeld, Germany: Springer.

Francc, a. A. C. e. s. C., Gouveia, T. B., Santos, P. C. F., Santana, C. A., & da Silva, F. Q. B. (2011). Motivation in software engineering: A systematic review update. In (Vol. 6, pp. 154–163).

Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 29*, 254–272 *6*.

Furui, S. (2005). 50 years of progress in speech and speaker recognition. *ECTI Transactions On Computer and Information Technology, 1*, 64–74.

Furui, S. (2009). Selected topics from 40 years of research on speech and speaker recognition. In *INTERSPEECH* (pp. 1–8).

Gabrea, M. (2011). Two microphones speech enhancement systems based on instrumental variable algorithm for speaker identification. In *24th Canadian conference on electrical and computer engineering (CCECE)* (pp. 569–572). Niagara Falls, ON, Canada: IEEE.

Gales, M., & Young, S. (1992). An improved approach to the hidden Markov model decomposition of speech and noise. In *Acoustics, speech, and signal processing, 1992. ICASSP-92., 1992 IEEE international conference on: 1* (pp. 233–236). IEEE.

Ganchev, T. (2011). *Contemporary methods for speech parameterization*. Springer.

Ghahabi, O., & Hernando, J. (2014). Deep belief networks for i-vector based speaker recognition. In *2014 IEEE International conference on acoustics, speech and signal processing (ICASSP)* (pp. 1700–1704).

Ghezaiel, W., Slimane, A. B., & Braiek, E. B. (2012). Usable speech assignment for speaker identification under co-channel situation. *International Journal of Computer Applications, 59*, 7–11.

Ghezaiel, W., Slimane, A. B., & Braiek, E. B. (2013). Improved EMD usable speech detection for co-channel speaker identification. In *International conference on nonlinear speech processing* (pp. 184–191). Springer.

Ghiurcau, M. V., Rusu, C., & Astola, J. (2011). A study of the effect of emotional state upon text-independent speaker identification. In *2011 IEEE International conference on acoustics, speech and signal processing (ICASSP)* (pp. 4944–4947). Prague, Czech Republic: IEEE.

Godin, K. W., Sadjadi, S. O., & Hansen, J. H. L. (2013). Impact of noise reduction and spectrum estimation on noise robust speaker identification. In *INTERSPEECH* (pp. 3656–3660).

Gong, C., Zhao, H., & Tao, Z. (2014). Speaker identification of whispered speech with perceptible mood. *Journal of Multimedia, 9*, 553–561 *6*.

Haigh, J. A., & Mason, J. S. (1993). Robust voice activity detection using cepstral features. In *1993 IEEE Region 10 conference on proceedings. computer, communication, control and power engineering (TENCON'93): 3* (pp. 321–324). IEEE.

Hanilcc, i. C., Kinnunen, T., Saeidi, R., Pohjalainen, J., Alku, P., & Ertas, F. (2013). Speaker identification from shouted speech: Analysis and compensation. In (pp. 8027–8031).

Hansen, J. H. L., & Hasan, T. (2015). Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Processing Magazine, 32*, 74–99.

Hu, Y., Wu, D., & Nucci, A. (2013). Fuzzy-clustering-based decision tree approach for large population speaker identification. *IEEE Transactions on Audio, Speech, and Language Processing, 21*, 762–774.

Hyon, S., Wang, H., Zhao, C., Wei, J., & Dang, J. (2012). A method of speaker identification based on phoneme mean F-ratio contribution. In *INTERSPEECH* (pp. 2670–2673).

Islam, M. R., & Rahman, M. F. (2009). Improvement of text dependent speaker identification system using neuro-genetic hybrid algorithm in office environmental conditions. *International Journal of Computer Science Issues, 1*, 42–48.

Jawarkar, N. P., Holambe, R. S., & Basu, T. K. (2012). Text-independent speaker identification in emotional environments: A classifier fusion approach. In *Frontiers in Computer Education* (pp. 569–576). Springer.

Jawarkar, N. P., Holambe, R. S., & Basu, T. K. (2015). Effect of nonlinear compression function on the performance of the speaker identification system under noisy conditions. In *Proceedings of the 2nd International Conference on Perception and Machine Intelligence* (pp. 137–144). Kolkata, West Bengal, India: ACM.

Jayanna, H., & Prasanna, S. M. (2009). Analysis, feature extraction, modeling and testing techniques for speaker recognition. *IETE Technical Review, 26*, 181–190.

Jiang, S., Frigui, H., & Calhoun, A. W. (2015). Speaker identification in medical simulation data using fisher vector representation. In *2015 IEEE 14th international conference on machine learning and applications (ICMLA)* (pp. 197–201). Miami, FL, USA: IEEE.

Jiang, T., Gao, B., & Han, J. (2009). *Speaker identification and verification from audio coded speech in matched and mismatched conditions* (pp. 2199–2204). IEEE.

Jin, Q., Toth, A. R., Schultz, T., & Black, A. W. (2009). *Speaker de-identification via voice transformation* (pp. 529–533). IEEE.

Jourani, R., Daoudi, K., Andre, O. R., & Aboutajdine, D. (2013). Combination of SVM and large margin GMM modeling for speaker identification. In *2013 Proceedings of the 21st european signal processing conference (EUSIPCO)* (pp. 1–5). Marrakech, Morocco: IEEE.

Justin, T., Struc, V., Dobrisek, S., Vesnicer, B., Ipsic, I., & Mihelic, F. (2015). In *Speaker de-identification using diphone recognition and speech synthesis: 4* (pp. 1–7). IEEE.

Kawakami, Y., Wang, L., Kai, A., & Nakagawa, S. (2014). Speaker identification by combining various vocal tract and vocal source features. In *International conference on text, speech, and dialogue* (pp. 382–389). Springer.

Kawakami, Y., Wang, L., & Nakagawa, S. (2013). Speaker identification using pseudo pitch synchronized phase information in noisy environments. In *2013 Asia-Pacific on signal and information processing association annual summit and conference (APSIPA)* (pp. 1–4). Kaohsiung, Taiwan: IEEE.

Keerio, A., Mitra, B. K., Birch, P., Young, R., & Chatwin, C. (2009). On preprocessing of speech signals. *International Journal of Signal Processing, 5*, 216–222.

Kekre, H. B., Athawale, A., & Desai, M. (2011). Speaker identification using row mean vector of spectrogram. In *Proceedings of the international conference and workshop on emerging trends in technology* (pp. 171–174).

Khan, S., Basu, J., & Bepari, M. S. (2012). Performance evaluation of PBDP based real-time speaker identification system with normal MFCC vs MFCC of LP residual features. In *Perception and machine intelligence* (pp. 358–366). Springer.

Kim, M.-J., Yang, I.-H., & Yu, H.-J. (2013). Histogram equalization using centroids of fuzzy C-Means of background speakers' utterances for speaker identification. In *SLSP'13 proceedings of the first international conference on statistical language and speech processing* (pp. 143–151). Tarragona, Spain: Springer.

Kinnunen, T. (2003). *Spectral features for automatic text-independent speaker recognition*. Joensuu, Finland: University of Joensuu.

Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech communication, 52*, 12–40.

Kitchenham, B., Pretorius, R., Budgen, D., Brereton, O. P., Turner, M., & Niazi, M. (2010). Systematic literature reviews in software engineering-a tertiary study. *Information and Software Technology, 52*, 792–805.

Kitchenham, B. A., & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical report EBSE-2007-01*: school of computer science and mathematics*. Keele University.

Kockmann, M., Burget, L., & Cernocký, J. H. (2011). Application of speaker-and language identification state-of-the-art techniques for emotion recognition. *Speech communication, 53*, 1172–1185.

Kundu, A., Das, D., & Bandyopadhyay, S. (2012). Speaker identification from film dialogues. In (pp. 1–4).

Lai, J.-Y., Wang, S.-L., Shi, X.-J., & Liew, A. W.-C. (2014). Sparse coding based lip texture representation for visual speaker identification. In *2014 19th international conference on digital signal processing (DSP)* (pp. 607–610). Hong Kong, China: IEEE.

Larcher, A., Lee, K. A., Ma, B., & Li, H. (2014). Text-dependent speaker verification: Classifiers, databases and RSR2015. *Speech communication, 60*, 56–77.

Lawson, A., Vabishchevich, P., Huggins, M., Ardis, P., Battles, B., & Stauffer, A. (2011). Survey and evaluation of acoustic features for speaker recognition. In *Acoustics, speech and signal processing (ICASSP), 2011 ieee international conference on* (pp. 5444–5447). IEEE.

LeCun, Y., Kavukcuoglu, K., & Farabet, C. (2010). Convolutional networks and applications in vision. In *ISCAS* (pp. 253–256).

Li, C.-H., Delbruck, T., & Liu, S.-C. (2012). Real-time speaker identification using the AEREAR2 event-based silicon cochlea. In *2012 IEEE international symposium on circuits and systems (ISCAS)* (pp. 1159–1162). Seoul, South Korea: IEEE.

Li, Q., & Huang, Y. (2011). An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions. *IEEE Transactions on Audio, Speech, and Language Processing, 19*, 1791–1801.

Li, Z., & Gao, Y. (2016). Acoustic feature extraction method for robust speaker identification. *Multimedia Tools Applications, 75*, 7391–7406.

Lippmann, R. P. (1989). Review of neural networks for speech recognition. *Neural computation, 1*, 1–38.

Liu, G., Lei, Y., & Hansen, J. H. L. (2012). Robust feature front-end for speaker identification. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4233–4236). Kyoto, Japan: IEEE.

Liu, T., & Guan, S. (2014). Factor analysis method for text-independent speaker identification. *Journal of Software, 9*, 2851–2860.

Lu, H., Brush, A. J. B., Priyantha, B., Karlson, A. K., & Liu, J. (2011). SpeakerSense: Energy efficient unobtrusive speaker identification on mobile phones. In *International conference on pervasive computing* (pp. 188–205). San Francisco, USA: Springer.

Luengo, I., Navas, E., Sainz, I. n. a., Saratxaga, I., Sanchez, J., & Odriozola, I. (2008). Text independent speaker identification in multilingual environments. In *Proceedings of the international conference on language resources and evaluation, LREC 2008*.

Lukic, Y., Vogt, C., Dürr, O., & Stadelmann, T. (2016). Speaker identification and clustering using convolutional neural networks. In *2016 IEEE 26th international workshop on machine learning for signal processing (MLSP)* (pp. 1–6).

Lyubimov, N., Nastasenko, M., Kotov, M., & Doroshin, D. (2014). Exploiting non-negative matrix factorization with linear constraints in noise-robust speaker identification. In *International conference on speech and computer* (pp. 200–208). Springer.

Ma, Z., & Leijon, A. (2011). Super-Dirichlet mixture models using differential line spectral frequencies for text-independent speaker identification. In *INTERSPEECH* (pp. 2360–2363).

Ma, Z., Yu, H., Tan, Z. H., & Guo, J. (2016). Text-independent speaker identification using the histogram transform model. *IEEE Access, 4*, 9733–9739.

Malegaonkar, A., & Ariyaeeinia, A. (2011). Performance evaluation in open-set speaker identification. In *European workshop on biometrics and identity management* (pp. 106–112). Springer.

Matějka, P., Glembek, O., Novotný, O., Plchot, O., Grézl, F., & Burget, L. ,&Cernocký,J. H. (2016). Analysis of DNN approaches to speaker identification. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5100–5104).

Matejka, P., Zhang, L., Ng, T., Mallidi, S. H., Glembek, O., & Ma, J. (2014). Neural network bottleneck features for language identification. In *Proc. of IEEE Odyssey* (pp. 299–304).

McLaren, M., Lei, Y., & Ferrer, L. (2015). Advances in deep neural network approaches to speaker recognition. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP): 6* (pp. 4814–4818).

McLaren, M., Scheffer, N., Graciarena, M., Ferrer, L., & Lei, Y. (2013). Improving speaker identification robustness to highly channel-degraded speech through multiple system fusion. In *2013 IEEE international conference on acoustics, speech and signal processing: 6* (pp. 6773–6777). Vancouver, BC, Canada: IEEE.

Meng, Y., Hu, Y., Zhang, H., & Wang, X. (2011). Speaker identification in time-sequential images based on movements of lips. In *2011 Eighth international conference on fuzzy systems and knowledge discovery (FSKD): 3* (pp. 1729–1733). Shanghai, China: IEEE.

Michalevsky, Y., Talmon, R., & Cohen, I. (2011). Speaker identification using diffusion maps. In *2011 19th European signal processing conference* (pp. 1299–1302). Barcelona, Spain: IEEE.

Mitra, V., McLaren, M., Franco, H., Graciarena, M., & Scheffer, N. (2013). Modulation features for noise robust speaker identification. In *INTERSPEECH* (pp. 3703–3707).

Mizobe, Y., Kurogi, S., Tsukazaki, T., & Nishida, T. (2012). Multistep speaker identification using Gibbs-distribution-based extended Bayesian inference for rejecting unregistered speaker. In *International conference on neural information processing* (pp. 247–255). Doha, Qatar: Springer.

Nagaraja, B. G., & Jayanna, H. S. (2012). Multilingual speaker identification with the constraint of limited data using multitaper MFCC. In *International conference on security in computer networks and distributed systems* (pp. 127–134). Springer.

Nagaraja, B. G., & Jayanna, H. S. (2013). Multilingual speaker identification by combining evidence from LPR and multitaper MFCC. *Journal of Intelligent Systems, 22*, 241–251.

Nakagawa, S., Wang, L., & Ohtsuka, S. (2012). Speaker identification and verification by combining MFCC and phase information. *IEEE Transactions on Audio, Speech & Language Processing, 20*, 1085–1095.

Nugraha, A. A., Yamamoto, K., & Nakagawa, S. (2014). Single-channel dereverberation by feature mapping using cascade neural networks for robust distant speaker identification and speech recognition. *EURASIP Journal on Audio, Speech, and Music Processing, 2014*, 1–31.

Ouamour, S., & Sayoud, H. (2013). Automatic speaker localization based on speaker identification-A smart room application. In *Fourth international conference on information and communication technology and accessibility (ICTA)* (pp. 1–5). Hammamet, Tunisia: IEEE.

Pal, A., Bose, S., Basak, G. K., & Mukhopadhyay, A. (2014). Speaker identification by aggregating Gaussian mixture models (GMMs) based on uncorrelated MFCC-derived features. *International Journal of Pattern Recognition and Artificial Intelligence, 28*.

Pathak, M. A., & Raj, B. (2013). Privacy-preserving speaker verification and identification using Gaussian mixture models. *IEEE Transactions on Audio, Speech, and Language Processing, 21*, 397–406.

Perner, P. (2010). *Case-Based reasoning on images and signals*. Springer Publishing Company, Incorporated.

Pierrehumbert, J. B. (1980). *The phonology and phonetics of english intonation*. Massachusetts Institute of Technology.

Plchot, O., Matsoukas, S., Matejka, P., Dehak, N., Ma, J. Z., & Cumani, S. (2013). Developing a speaker identification system for the DARPA RATS project. In *2013 IEEE International conference on acoustics, speech and signal processing (ICASSP)* (pp. 6768–6772). Vancouver, BC, Canada: IEEE.

Pobar, M., & Ipsic, I. (2014). Online speaker de-identification using voice transformation. In *2014 37th International convention on information and communication technology, electronics and microelectronics (mipro)* (pp. 1264–1267). opatija, croatia: IEEE.

Poignant, J., Besacier, L., & Quénot, G. (2015). Unsupervised speaker identification in TV broadcast based on written names. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23*, 57–68.

Prakash, C., & Gangashetty, S. V. (2011). Fourier-bessel based cepstral coefficient features for text-independent speaker identification. In *5 th Indian international conference on artificial intelligence (IICAI-11)* (pp. 913–930).

Prasad, A., Periyasamy, V., & Ghosh, P. K. (2015). Estimation of the invariant and variant characteristics in speech articulation and its application to speaker identification. In *2015 IEEE International conference on acoustics, speech and signal processing (ICASSP)* (pp. 4265–4269). Brisbane, QLD, Australia: IEEE.

Prasad, S., Tan, Z.-H., & Prasad, R. (2013). Multi-frame rate based multiple-model training for robust speaker identification of disguised voice. In *2013 16th international symposium on wireless personal multimedia communications (WPMC)* (pp. 1–4). Atlantic City, NJ, USA: IEEE.

Qi, J., Wang, D., Xu, J., & Tejedor Noguerales, J. (2013). Bottleneck features based on gammatone frequency cepstral coefficients. *Interspeech*. International Speech Communication Association.

Qi, P., & Wang, L. (2011). Experiments of GMM based speaker identification. In *2011 8th International conference onubiquitous robots and ambient intelligence (URAI)* (pp. 26–31). Incheon, South Korea: IEEE.

Rajesh, R., Ganesh, K., Koh, S. C. L., Singh, N., Khan, R. A., & Shree, R. (2012). International conference on modelling optimization and computingapplications of speaker recognition. *Procedia Engineering, 38*, 3122–3126.

Ram, i,r. J., Segura, J. e. C., Ben, i,t. C., De La Torre, A., & Rubio, A. (2004). Efficient voice activity detection algorithms using long-term speech information. *Speech Communication, 42*, 271–287.

Ramachandran, R. P., Polikar, R., Dahm, K. D., & Shetty, S. S. (2012). Open-ended design and performance evaluation of a biometric speaker identification system. In *2012 IEEE International symposium on circuits and systems (ISCAS)* (pp. 2697–2700). Seoul, South Korea: IEEE.

Rao, K. S., & Sarkar, S. (2014a). Robust speaker verification: A review. In *Robust speaker recognition in noisy environments* (pp. 13–27). Springer.

Raval, K., Ramachandran, R. P., Shetty, S. S., & Smolenski, B. Y. (2012). Feature and signal enhancement for robust speaker identification of g. 729 decoded speech. In *ICONIP'12 Proceedings of the 19th international conference on neural information processing* (pp. 345–352). Doha, Qatar: Springer.

Rekik, A., Ben-Hamadou, A., & Mahdi, W. (2015). Unified system for visual speech recognition and speaker identification. In *ACIVS 2015 proceedings of the 16th international conference on advanced concepts for intelligent vision systems: 9386* (pp. 381–390). Catania, Italy: Springer.

Revathi, A., & Venkataramani, Y. (2009). Text independent composite speaker identification/verification using multiple features. In *2009 WRI World congress on computer science and information engineering: 7* (pp. 257–261).

Reynolds, D. (2002). An overview of automatic speaker recognition. In *Proceedings of the international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4072–4075).

Richardson, F., Reynolds, D., & Dehak, N. (2015a). Deep neural network approaches to speaker and language recognition. *IEEE Signal Processing Letters, 22*, 1671–1675.

Richardson, F., Reynolds, D., & Dehak, N. (2015b). A unified deep neural network for speaker and language recognition. *arXiv, 6*.

Rossi, M., Amft, O., & Tröster, G. (2012). Collaborative personal speaker identification: A generalized approach. *Pervasive and Mobile Computing, 8*, 415–428.

Sadjadi, S. O., & Hansen, J. H. L. (2011). Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions. In *2011 IEEE International conference on acoustics, speech and signal processing (ICASSP)* (pp. 5448–5451). Prague, Czech Republic: IEEE.

Sadjadi, S. O., & Hansen, J. H. L. (2013). Robust front-end processing for speaker identification over extremely degraded communication channels. In *2013 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 7214–7218). Vancouver, BC, Canada: IEEE.

Sadjadi, S. O., & Hansen, J. H. L. (2015). Mean Hilbert envelope coefficients (MHEC) for robust speaker and language identification. *Speech communication, 72*, 138–148.

Saeidi, R., Hurmalainen, A., Virtanen, T., & van Leeuwen, D. A. (2012). Exemplar-based sparse representation and sparse discrimination for noise robust speaker identification. In *Proceedings of the speaker and language recognition workshop odyssey* (pp. 248–255).

Safavi, S., Hanani, A., Russell, M., Jancovic, P., & Carey, M. J. (2012). Contrasting the effects of different frequency bands on speaker and accent identification. *IEEE Signal Processing Letters, 19*, 829–832.

Sahidullah, M., Chakroborty, S., & Saha, G. (2011). Improving performance of speaker identification system using complementary information fusion. In *Proceedings of 17th international conference on advanced computing and communications* (pp. 182–187). ArXiv.

Sahidullah, M., & Saha, G. (2011). In search of autocorrelation based vocal cord cues for speaker identification. In *Proceedings of 2nd international conference on RF & signal processing systems - RSPS 2010* (pp. 5–11).

Salapa, K., Trawińska, A., Roterman, I., & Tadeusiewicz, R. (2014). Speaker identification based on artificial neural networks. Case study: The Polish vowel (pilot study). *Bio-Algorithms and Med-Systems, 10*, 91–99.

Saquib, Z., Salam, N., Nair, R. P., Pandey, N., & Joshi, A. (2010a). A survey on automatic speaker recognition systems. In T.-h. Kim, S. K. Pal, W. I. Grosky, N. Pissinou, T. K. Shih, & D. Ślęzak (Eds.), *Signal processing and multimedia: international conferences, SIP and MulGraB 2010, held as part of the future generation information technology conference, FGIT 2010, Jeju Island, Korea, December 13-15, 2010. proceedings* (pp. 134–145). Berlin, Heidelberg: Springer Berlin Heidelberg.

Saquib, Z., Salam, N., Nair, R. P., Pandey, N., & Joshi, A. (2010b). A survey on automatic speaker recognition systems. *Signal Processing and Multimedia*, 134–145.

Sarangi, S. K., & Saha, G. (2012). A novel approach in feature level for robust text-independent speaker identification system. In *2012 4th international conference on intelligent human computer interaction (IHCI)* (pp. 1–5). Kharagpur, India: IEEE.

Sarkar, A. K., & Umesh, S. (2011). Eigen-voice based anchor modeling system for speaker identification using MLLR super-vector. In *INTERSPEECH* (pp. 2357–2360).

Sarkar, A. K., Umesh, S., & Bonastre, J.-F. c. o. (2012). Computationally efficient speaker identification using fast-MLLR based anchor modeling. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4357–4360). Kyoto, Japan: IEEE.

Sarma, M., & Sarma, K. K. (2013a). Speaker identification model for Assamese language using a neural framework. In *The 2013 international joint conference on neural networks (IJCNN)* (pp. 1–7). Dallas, TX, USA: IEEE.

Sarma, M., & Sarma, K. K. (2013b). Vowel phoneme segmentation for speaker identification using an ANN-based framework. *Journal of Intelligent Systems, 22*, 111–130.

Schmidt, L., Sharifi, M., & Moreno, I. L. (2014). Large-scale speaker identification. In *2014 IEEE International conference on acoustics, speech and signal processing (ICASSP)* (pp. 1650–1654). Florence, Italy: IEEE.

Sen, N., & Basu, T. K. (2011a). Features extracted using frequency-time analysis approach from Nyquist filter bank and Gaussian filter bank for text-independent speaker identification. In *European workshop on biometrics and identity management* (pp. 125–136). Brandenburg, Germany: Springer.

Sen, N., & Basu, T. K. (2011b). Features Extracted Using Frequency-Time Analysis Approach from Nyquist Filter Bank and Gaussian Filter Bank for Text-Independent Speaker Identification. In (pp. 125–136).

Sen, N., & Basu, T. K. (2012). A critical comparison between GMM classifier and polynomial classifier for text-independent speaker identification. In *Frontiers in computer education* (pp. 545–550). Springer.

Shahamiri, S. R., Kadir, W. M. N. W., Ibrahim, S., & Hashim, S. Z. B. (2012). Artificial neural networks as multi-networks automated test oracle. *Automated Software Engineering, 19*, 303–334.

Shahamiri, S. R., & Salim, S. S. B. (2014a). Artificial neural networks as speech recognisers for dysarthric speech: Identifying the best-performing set of MFCC parameters and studying a speaker-independent approach. *Advanced Engineering Informatics, 28*, 102–110.

Shahamiri, S. R., & Salim, S. S. B. (2014b). A multi-views multi-learners approach towards dysarthric speech recognition using multi-nets artificial neural networks. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 22*, 1053–1063.

Shahamiri, S. R., & Salim, S. S. B. (2014c). Real-time frequency-based noise-robust automatic speech recognition using multi-nets artificial neural networks: A multi-views multi-learners approach. *Neurocomputing, 129*, 199–207.

Shahin, I. (2013). Speaker identification in emotional talking environments based on CSPHMM2s. *Engineering Applications of Artificial Intelligence, 26*, 1652–1659.

Shih, P.-Y., Lin, P.-C., Wang, J.-F., & Lin, Y.-N. (2011). Robust several-speaker speech recognition with highly dependable online speaker adaptation and identification. *Journal of network and computer applications, 34*, 1459–1467.

Sidorov, M., Schmitt, A., Zablotskiy, S., & Minker, W. (2013). Survey of automated speaker identification methods. In *2013 9th international conference on intelligent environments (IE)* (pp. 236–239). Athens, Greece: IEEE.

Singh, P., Laxmi, V., & Gaur, M. S. (2012). Speaker identification using optimal lip biometrics. In *2012 5th IAPR international conference on biometrics (ICB* (pp. 472–477). New Delhi, India: IEEE.

Sreenivasa Rao, K., & Sarkar, S. (2014). *Robust speaker recognition in noisy environments*. Springer International Publishing.

Srinivas, V., Rani, C. S., & Madhu, T. (2014). Neural network based classification for speaker identification. *International Journal of Signal Processing, Image Processing and Pattern Recognition, 7*, 109–120.

Srinivasan, R., Ming, J., & Crookes, D. (2012). Single-channel speaker-pair identification: A new approach based on automatic frame selection. In (pp. 4369–4372).

Srivastava, S., Bhardwaj, S., Bhandari, A., Gupta, K., Bahl, H., & Gupta, J. R. P. (2013). Wavelet packet based mel frequency cepstral features for text independent speaker identification. In *Intelligent informatics* (pp. 237–247). Springer.

Stapic, Z., Lopez, E. G., Cabot, A. G., de Marcos Ortega, L., & Strahonja, V. (2012). Performing systematic literature review in software engineering. In S. Young, E. G. Gales Mark, T. Hain, D. Kershaw, X. (Andrew) Liu, & G. Moore (Eds.). *HTK book (for HTK version 3.4)*: 2. Camb. Univ. Eng. Dep.

Taghia, J., Ma, Z., & Leijon, A. (2013). On von-Mises Fisher mixture model in text-independent speaker identification. In *INTERSPEECH* (pp. 2499–2503).

Tanprasert, C., & Achariyakulporn, V. (2000). Comparative study of GMM, DTW, and ANN on Thai speaker identification system. *Sixth international conference on spoken language processing, ICSLP 2000 / INTERSPEECH 2000*. Beijing, China: ISCA.

Tirumala, S. S., & Shahamiri, S. R. (2016). A review on deep learning approaches in

speaker identification. In *Proceedings of the 8th international conference on signal processing systems* (pp. 142–147). ACM.
Togneri, R., & Pullella, D. (2011). An overview of speaker identification: accuracy and robustness issues. *IEEE Circuits and Systems Magazine, 11*, 23–61.
Tomar, V. S., & Rose, R. C. (2013). Efficient manifold learning for speech recognition using locality sensitive hashing. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 6995–6999). IEEE.
Trabelsi, I., & Ayed, D. B. (2014). A multi level data fusion approach for speaker identification on telephone speech. *International Journal of Signal Processing, Image Processing and Pattern Recognition, 6*.
Vandyke, D., Wagner, M., & Goecke, R. (2013). Voice source waveforms for utterance level speaker identification using support vector machines. In *Information Technology in Asia (CITA), 2013 8th International Conference on* (pp. 1–7). Kota Samarahan, Malaysia: IEEE.
Vannicola, C. M., Smolenski, B. Y., Battles, B., & Ardis, P. A. (2011). Mitigation of reverberation on speaker identification via homomorphic filtering of the linear prediction residual. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5512–5515). Prague, Czech Republic: IEEE.
Vasudev, D., & K, A. B. K. (2014). Speaker Identification using FBCC in Malayalam language. In *2014 international conference on advances in computing, communications and informatics (ICACCI* (pp. 1759–1763). New Delhi, India: IEEE.
Verma, G. K. (2011). Multi-feature fusion for closed set text independent speaker identification. In *International conference on information intelligence, systems, technology and management* (pp. 170–179). Springer.
Volfin, I., & Cohen, I. (2013). Dominant speaker identification for multipoint videoconferencing. *Computer Speech & Language, 27*, 895–910.
Wang, J.-C., Chin, Y.-H., Hsieh, W.-C., Lin, C.-H., Chen, Y.-R., & Siahaan, E. (2015). Speaker identification with whispered speech for the access control system. *IEEE Transactions on Automation Science and Engineering, 12*, 1191–1199 *6*.
Wang, J.-F., Peng, J.-S., Wang, J.-C., Lin, P.-C., & Kuan, T.-W. (2011). Hardware/software co-design for fast-trainable speaker identification system based on SMO. In *2011 IEEE international conference on systems, man, and cybernetics (SMC)* (pp. 1621–1625). Anchorage, AK, USA: IEEE.
Wang, J., & Johnson, M. T. (2014). Physiologically-motivated feature extraction for speaker identification. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 1690–1694). Florence, Italy: IEEE.
Wang, L., Zhang, Z., Kai, A., & Kishi, Y. (2012a). Hands-free speaker identification based on spectral subtraction using a multi-channel least mean square approach. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 7224–7228). Vancouver, BC, Canada: IEEE.
Wang, L., Zhang, Z., Kai, A., & Kishi, Y. (2012a). Distant-talking speaker identification using a reverberation model with various artificial room impulse responses. In *2012 Asia-Pacific signal & information processing association annual summit and conference (APSIPA ASC)* (pp. 1–4). Hollywood, CA, USA: IEEE.
Wang, Y., Tang, F., & Zheng, J. (2012b). Robust text-independent speaker identification in a time-varying noisy environment. *Journal of Software, 7*, 1975–1980.
Wu, J.-D., & Lin, B.-F. (2009). Speaker identification using discrete wavelet packet transform technique with irregular decomposition. *Expert Systems with Applications, 36*, 3136–3143 *6*.
Wu, J.-D., & Tsai, Y.-J. (2011). Speaker identification system using empirical mode decomposition and an artificial neural network. *Expert Systems with Applications, 38*, 6112–6117 *6*.
Xie, J., Xu, L., & Chen, E. (2012). Image denoising and inpainting with deep neural networks. In *Advances in neural information processing systems* (pp. 341–349).
Xing, Y., Li, H., & Tan, P. (2012). Hierarchical fuzzy speaker identification based on FCM and FSVM. In *2012 9th international conference on fuzzy systems and knowledge discovery (FSKD)* (pp. 311–315). Sichuan, China: IEEE.
Yamada, T., Wang, L., & Kai, A. (2013). Improvement of distant-talking speaker identification using bottleneck features of DNN. In *INTERSPEECH* (pp. 3661–3664). Lyon, France: ISCA.
Yang, I. L. H., Kim, M.-S., So, B.-M., Kim, M.-J., & Yu, H.-J. (2012). Robust speaker identification using ensembles of kernel principal component analysis. In *7th international conference on hybrid artificial intelligent systems* (pp. 71–78). Salamanca, Spain: Springer.
Yang, Y., Chen, L., & Wang, W. (2011). Emotional speaker identification by humans and machines. In *CCBR'11 proceedings of the 6th chinese conference on biometric recognition* (pp. 167–173). Beijing, China: Springer.
Yang, Y., & Liu, J. (2014). Dereverberation for speaker identification in meeting. In *International conference on human-computer interaction* (pp. 594–599). Springer.
Yu, H., Ma, Z., Li, M., & Guo, J. (2014). Histogram transform model using MFCC features for text-independent speaker identification. In *2014 48th Asilomar conference on signals, systems and computers: 6* (pp. 500–504).
Zao, L., & Coelho, R. (2011). Colored noise based multicondition training technique for robust speaker identification. *IEEE Signal Processing Letters, 18*, 675–678.
Zhang, X., Zhang, H., & Gao, G. (2014a). Missing feature reconstruction methods for robust speaker identification. In *2014 Proceedings of the 22nd european signal processing conference (EUSIPCO)* (pp. 1482–1486). Lisbon, Portugal: IEEE.
Zhang, X. y., Bai, J., & Liang, W. z. (2006). The speech recognition system based on bark wavelet MFCC. *2006 8th international conference on signal processing*: 1. Beijing, China: IEEE.
Zhang, Z., Wang, L., & Kai, A. (2014b). Distant-talking speaker identification by generalized spectral subtraction-based dereverberation and its efficient computation. *EURASIP Journal on Audio, Speech, and Music Processing, 2014*(6), 1–12.
Zhang, Z., Wang, L., Kai, A., Yamada, T., Li, W., & Iwahashi, M. (2015). Deep neural network-based bottleneck feature and denoising autoencoder-based dereverberation for distant-talking speaker identification. *EURASIP Journal on Audio, Speech, and Music Processing, 2015*, 1–13.
Zhao, C., Wang, H., Hyon, S., Wei, J., & Dang, J. (2012). Efficient feature extraction of speaker identification using phoneme mean F-ratio for Chinese. In *2012 8th international symposium on chinese spoken language processing (ISCSLP)* (pp. 345–348). Kowloon, China: IEEE.
Zhao, G., & Pietikäinen, M. (2013). Visual speaker identification with spatiotemporal directional features. In *International conference image analysis and recognition* (pp. 1–10). Springer.
Zhao, X., Shao, Y., & Wang, D. (2011). Robust speaker identification using a CASA front-end. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5468–5471). Prague, Czech Republic: IEEE.
Zhao, X., Shao, Y., & Wang, D. (2012). CASA-based robust speaker identification. *IEEE Transactions on Audio, Speech, and Language Processing, 20*, 1608–1616.
Zhao, X., & Wang, D. (2013). Analyzing noise robustness of MFCC and GFCC features in speaker identification. In *2013 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 7204–7208). Vancouver, BC, Canada: IEEE.
Zhao, X., Wang, Y., & Wang, D. (2014). Robust speaker identification in noisy and reverberant conditions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22*, 836–845.
Zhao, X., Wang, Y., & Wang, D. (2015). Cochannel speaker identification in anechoic and reverberant conditions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23*, 1727–1736 *6*.