Digital Signal Processing 110 (2021) 102951

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

A survey of speech emotion recognition in natural environment

Md. Shah Fahad [a,*], Ashish Ranjan [a], Jainath Yadav [b], Akshay Deepak [a]

[a] Department of Computer Science, National Institute of Technology Patna, India
[b] Department of Computer Science, Central University of South Bihar, Patna, India

ARTICLE INFO

Article history:
Available online 30 December 2020

Keywords:
Language-independent
Natural-environment
Noisy-environment
Speech emotion recognition (SER)
Speaker-independent
Text-independent

ABSTRACT

While speech emotion recognition (SER) has been an active research field since the last three decades, the techniques that deal with the natural environment have only emerged in the last decade. These techniques have reduced the mismatch in the distribution of the training and testing data, which occurs due to the difference in speakers, texts, languages, and recording environments between the training and testing datasets. Although a few good surveys exist for SER, they either don't cover all aspects of SER in natural environments or don't discuss the specifics in detail. This survey focuses on SER in a natural environment, discussing SER techniques for natural environment along with their advantages and disadvantages in terms of speaker, text, language, and recording environments. In the recent past, the deep learning techniques have become very popular due to minimal speech processing and enhanced accuracy. Special attention has been given to deep-learning techniques and the related issues in this survey. Recent databases, features, and feature selection algorithms for SER, which have not been discussed in the existing surveys and can be promising for SER in a natural environment, have also been discussed in this paper.

© 2020 Published by Elsevier Inc.

1. Introduction

Emotion plays an important role in real-life applications. The emotion of a person can be identified by various sources of information like speech, transcript, facial expression, brain signal (EEG), and a combination of two or more of these (called multi-modal emotion recognition). Among these sources, speech is arguably the easiest to acquire. Whether it is the physical movement of the speakers or visual occlusions due to glass, mustache, beard, etc., the speech attributes are little or not affected by these, as compared to facial expressions, which can be significantly affected. The speech (acoustic) features for emotion recognition are almost similar in all languages and the same classification model can be used across languages [1]. Even though SER has been carried out for most of the major languages, the same model can be used for other languages with acceptable accuracy. However, this is not true for linguistic features. Each language requires specific database. Compare this with the linguistic features where a language-specific model is required, which is often a hindrance for less popular or regional languages for whom labeled emotion data is not yet present. Even if labeled data is available, a language-specific model needs to be constructed, which is an obvious overhead.

Automatic SER is used in several applications. It enhances human-computer interaction (HCI) systems, such as interactive movies [2], storytelling, and E-tutoring applications [3], and retrieval and indexing of video/audio files [4]. Speech-based emotion recognition system assists in improving the quality of service of the call attendants at call centers [5]. Automatic emotion detection could be helpful in psychological treatments as used in [6], [7], [8]. It can also be useful in the case of surveillance systems [9]. Modern speech-based systems are designed largely using neutral speech. Here, the components of emotion can be used as an add-on to improve the accuracy in the practical applications. Nowadays, the voice-assisted search engines have become very popular. It would be beneficial for the on-line market to update a web page dynamically according to the user's emotion (detected through speech).

Automatic speech emotion recognition (SER) is achieved by the development of methodologies based on digital signal processing and machine learning. The journey of research in this field is three decades-long; still, the results are not good enough to be applied in natural environments with high accuracy. There is a multitude of information present in the speech signal. A speech signal contains lexical contents (what has been spoken), speaker (who is the speaker), emotions (how it has been spoken), and language (in which language it has been spoken). If one has to recognize particular information in speech, then ideally the effect of other

* Corresponding author.
E-mail addresses: shah.cse16@nitp.ac.in (Md. Shah Fahad), ashish.cse16@nitp.ac.in (A. Ranjan), jainath@cub.ac.in (J. Yadav), akshayd@nitp.ac.in (A. Deepak).

https://doi.org/10.1016/j.dsp.2020.102951
1051-2004/© 2020 Published by Elsevier Inc.

Table 1
Comparison of the proposed survey with the existing surveys in speech emotion recognition.

Survey paper	Year	Database	Features	Feature selection	Classifiers	Deep-learning	Speaker independent SER	Text independent SER	Language independent SER	SER in uncontrolled environment
Schuller et al. [10]	2011	✓	✓	✓	✓	X	✓	X	X	✓
Ayadi et al. [11]	2011	✓	✓	X	✓	X	X	X	X	X
Koolagudi and Rao [12]	2012	✓	✓	X	✓	X	X	X	X	X
Anagnostopoulos et al. [13]	2015	X	✓	✓	X	X	X	X	X	X
Swain et al. [14]	2018	✓	✓	X	✓	X	X	X	X	X
Mustafa et al. [15]	2018	✓	✓	X	✓	X	X	X	X	X
Schuller et al. [16]	2018	✓	✓	X	✓	✓	X	X	X	X
Akçay et al. [17]	2020	✓	✓	X	✓	✓	X	X	X	X
This Survey		✓	✓	✓	✓	✓	✓	✓	✓	✓

information should be nullified. For example, if one has to recognize emotion from speech, then the effect of the speaker, lexical content, and language should ideally be nullified to generalize the SER system. This is the primary reason why automatic SER systems don't work very well for real-life applications. This problem occurs due to mismatch of speaker, text, language, and culture – collectively referred to as 'environment' – in the training and testing data. As a result, the accuracy significantly decreases in the case of real-life applications or 'natural environment'. Here, 'natural' refers to the variation of speakers, text, language, culture, surroundings, etc., within and across the development and deployment environments of SER systems.

To deal with this challenge, researchers have focused on independent environments for training and testing. Specifically, this means speaker independence [18–20] (using different speakers for training and testing), text independence [21–24] (using different transcripts for training and testing), and language independence [25–27] (using different languages for training and testing, i.e., cross-corpus). Among these, cross-corpus experiments [25–27] are more popular because the information related to speaker, text, and language are, by default, different in each corpus. Various methods were proposed to deal with these issues. This survey highlights those techniques along with their advantages and disadvantages. Another type of mismatch is related to the variation in recording environments. This mismatch is created by different types of noises, compression of data, variation in microphone-distance, and channel mismatch. To deal with this mismatch problem of recording environments, recently, various noise compensation techniques and other approaches have been proposed. In this survey, the comprehensive study of those techniques is provided with their advantages and disadvantages.

There exist some good surveys [10,11,13–17,28] available in SER. Table 1 compares the existing surveys with this paper with respect to different aspects of SER. The existing surveys discuss the essential components of SER: (1) database, (2) features, (3) feature selection, and (4) classifiers. Schuller et al. [10] presented an end to end survey from databases to evaluation techniques. In their paper, the task of emotion recognition has been discussed both as a classification and regression problem. The databases, features, and feature selection techniques have been discussed in detail. Their survey explains various techniques to evaluate SER models. Their paper also highlights the research issues related to the robustness of SER in the noisy and cross-corpus environments but does not discuss the methods.

Koolagudi and Rao [28] presented a survey of databases, features, and classifiers. They categorized the databases into acted, elicited, and natural categories. They discussed their advantages and disadvantages. They categorized the features into excitation, vocal-tract, and prosodic features. They studied the effect of the combination of different features. Different classification methods and their combinations were also discussed for SER. They did not discuss the current emerging deep-learning techniques and issues for SER. They did not discuss the issues and techniques to deal with the natural environment.

Ayadi et al. [11] described the databases, features, and classifiers for SER. Their survey briefly discussed databases. A different and useful aspect of their survey is the categorization of features with respect to time (local vs. global) and source (continuous, qualitative, spectral, and TEO-based). A discussion on the combination of speech features with linguistic, face, video, and discourse features to enhance emotion recognition accuracy was also included in the survey. They did not focus on deep-learning techniques and issues related to SER. They also did not discuss the current focus such as the natural environment of SER.

Anagnostopoulos et al. [13] presented a survey on research in SER from 2000 to 2011, in which they presented an overall pipeline for SER. They highlighted the use of speech (acoustic) features in combination with linguistic and non-linguistic vocalization features (laughs, cries, sighs, yawns, etc.) for emotion recognition. The importance of the combination of classifiers, ensemble learning, and voting technique was discussed for emotion recognition. The feature selection algorithms were also discussed to some extent. They did not cover the current focus of SER such as the natural environment.

Swain et al. [14] reviewed the databases, features, and classifier techniques for SER. They categorized features as prosody, excitation, vocal-tract, and a fusion of one or more of these features for emotion recognition. Their paper also highlighted the usefulness of deep learning, hybrid, and fusion techniques for emotion classification. Mustafa et al. [15] reviewed the articles from 2006 to 2017. They pointed out and discussed the current focus of SER being cross-lingual and real-time SER. However, they did not discuss the techniques that handle the issues of cross-lingual and real-time SER.

Schuller et al. [16] presented a short survey in which they discussed the traditional and current approach of SER system. They focused on end-to-end SER system and discussed the current challenge of SER that comes through different languages and cultures. However, they did not discuss the issues and techniques related to end-to-end system as well as the natural environment.

Akçay et al. [17] presented a survey on SER that covers almost all the areas of SER, the emotion models, databases, features, supporting modalities and classifiers. In their paper, they discussed deep-learning classifiers and deep-learning-based enhancement techniques such as auto-encoder, multi-tasking, adversarial training, attention to detail. But they did not discuss the issues

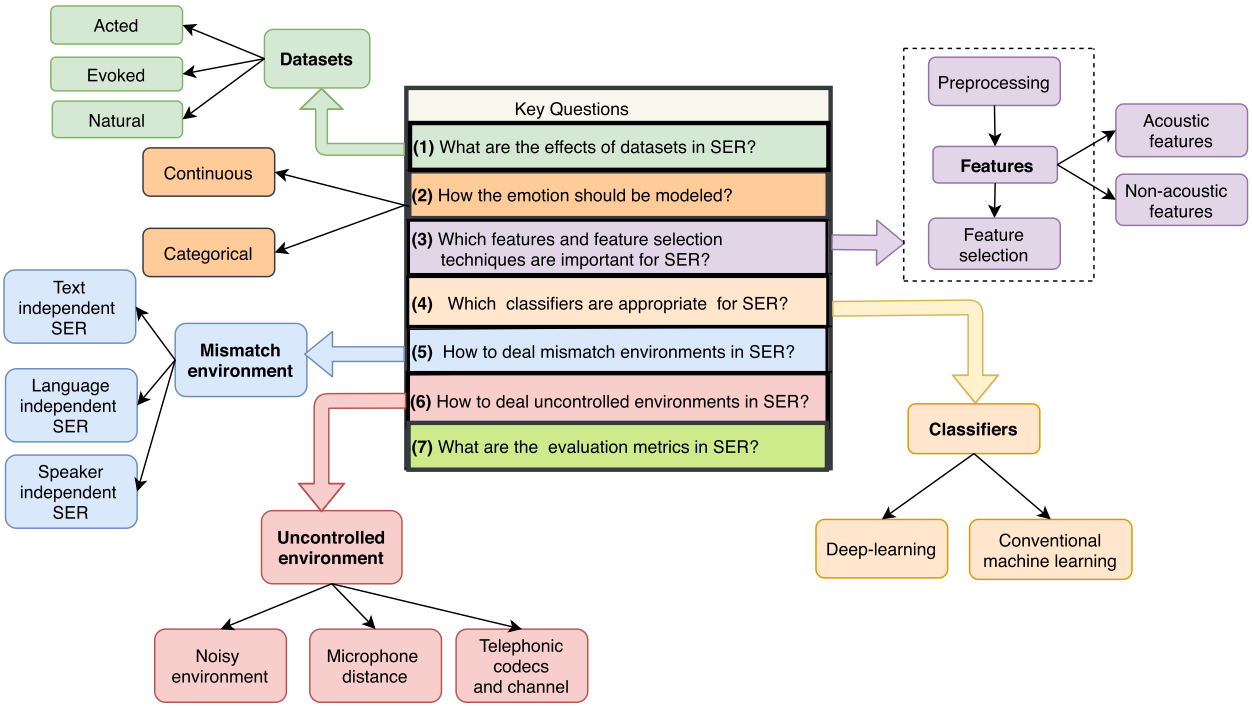

Fig. 1. Survey overview.

with respect to SER. They focused on deep-learning techniques but did not discuss the techniques related to the natural environment.

The existing survey papers and the current survey have been categorized with respect to the following attributes in Table 1: (1) database, (2) features, (3) feature selection, (4) classifiers, (5) deep-learning, (6) speaker-independent SER, (7) text-independent SER, (8) language-independent SER, and (9) SER in an uncontrolled environment.

The existing surveys discuss popular databases but miss some of the recently developed challenging natural databases [29–32]. This survey highlights the cause and potential solution for developing SER model from natural databases. The existing surveys discuss features that were commonly used for SER have missed out on the recently proposed wavelet-based, non-linear, and other features (discussed in Section 4.2.1). This survey includes those features and suggests which features are suitable for which category of emotion. There are a few surveys that discuss the importance of feature selection for speech emotion recognition. This survey discusses feature selection algorithms used for emotion recognition that have been adopted from other applications as well as those that have been specifically designed for emotion recognition. Most of the existing surveys discuss conventional classifiers like support vector machine (SVM), hidden Markov model (HMM), etc. Some of the surveys discussed the deep-learning methods but did not discuss their issues related to speech emotion recognition. Due to the grand success of deep-learning methods, which minimize the requirements of speech processing, they demand to be discussed in detail. In this survey, the deep-learning methods and their issues are discussed in detail for speech emotion recognition. Some of the surveys highlighted the current focus of SER, such as cross-lingual SER and SER in a noisy environment. However this survey discusses the current focus of research in detail with respect to issues and techniques, such as mismatched environments (Speaker-Independent SER, Text-Independent SER, and Language-Independent SER) and uncontrollable elements (noisy, microphone-distance, and telephonic codec).

Fig. 1 gives the overview of this survey along with the key questions being addressed. The sub-aspects of each of these questions are also shown in the figure. These seven key questions and their sub-aspects have been discussed in Sections 2-6.

The complete pipeline of a typical SER is presented in Fig. 2. The whole pipeline is divided into three parts: features (described in Section 4), model (described in Section 5), and testing. The feature part is common for both the training and testing phases. There are three steps in the feature part: preprocessing, feature extraction, and feature selection. Preprocessing is used to denoise the speech and to find salient segments. Feature extraction is a necessary step to determine the discriminative features corresponding to different emotions. The speech signal is non-stationary, therefore, the frame-level (local) features are extracted first and then the statistical descriptors are applied to obtain the global features. The resulting features are further normalized for speaker-independent emotion recognition using the statistics of the speakers in the training set. The feature selection strategy prevents the model from the curse of dimensionality by choosing non-redundant and relevant features. The model part either uses conventional or deep-learning-based models. The deep-learning-based models require minimal signal processing as the global features are learned by the model itself. The hierarchical classifier, which has been shown to produce good results for conventional models [33–35], can also be applied for deep-learning-based models. The distribution of the testing data is different than the training data due to the mismatch of speaker, text, and language. Therefore, an adaptation of existing model through testing data is required before the testing phase. The dotted block shows steps that are optional in general but are a must in the case of a natural environment to achieve better accuracy for emotion recognition.

The rest of the paper is organized in the following manner: In Section 2, the different types of databases are discussed along with their advantages and disadvantages. Section 3 describes how an emotion can be modeled. Section 4 discusses preprocessing, features, and feature selection algorithms for speech emotion recognition. In Section 5, different classifier techniques with their pros and cons are discussed in detail. The deep-learning techniques, their issues, and related solutions are also discussed in detail. Section 6 discusses the techniques related to the mismatch of envi-

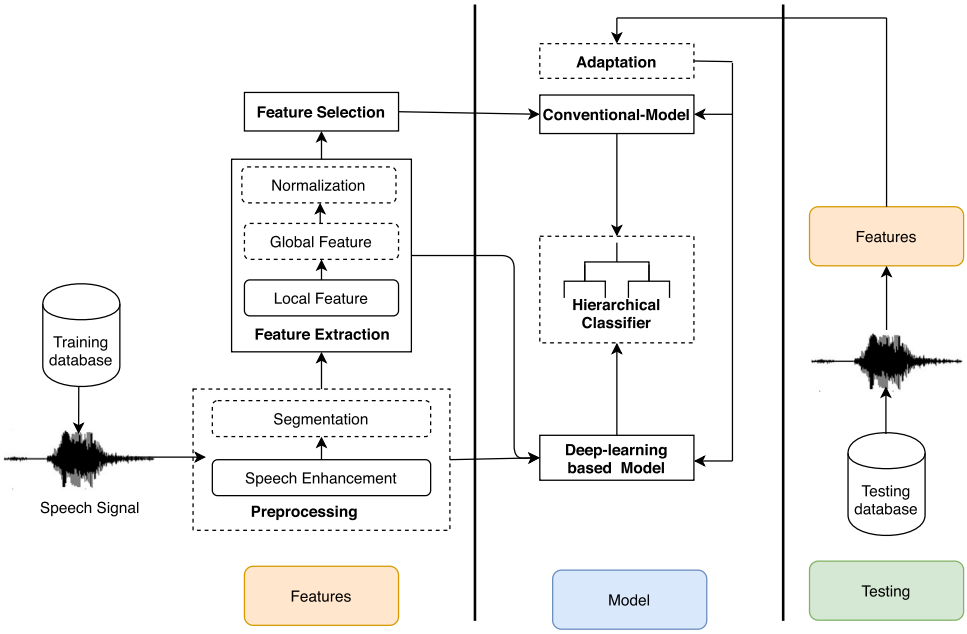

Fig. 2. The complete pipeline of a typical SER system is presented. The dotted line indicates steps specific to SER in a natural environment. The 'Feature' part is required in both the training and testing phases.

Table 2
Pros and Cons of different types of databases.

Types of Database	Pros	Cons
Acted	• All desired emotions are available. • Standardized and predefined speaker and text. • Easy to compare with existing works.	• Large cost in database construction. • There is a lack of context in the utterances. • Read-out emotions; absence of natural emotions.
Evoked	• Close to natural and context is present.	• Context is artificial. • All emotions may not be present. • Imbalanced data.
Natural	• Real emotions • Such databases are abundant in nature. • Low cost in database construction. • Contextual information is present.	• Data imbalanced with respect to emotions and utterance lengths. • All emotions may not be present. • Ethical issues in using a database. • Presence of background noise. • Overlapping of multiple emotions in an utterance and utterances themselves.

ronments due to human factors (speaker, text, and language) and recording environments (noise, microphone, codec mismatch, etc.). Section 7 discusses the evaluation metrics that are used to evaluate the model. Section 8 concludes the survey with pointers to further scope of work for SER in a natural environment.

2. Databases

One of the main issues in automatic speech emotion recognition is to collect appropriate data for training the model. Based on how emotional speech is generated, the emotional speech databases are categorized into three groups [12]: (1) acted, (2) evoked, and (3) natural. Table 2 summarizes the pros and cons of each type of database. While the acted and evoked databases have been discussed briefly on account of being non-natural, the natural databases have been discussed in much more detail being suitable for SER in a natural environment.

1. Acted database: Acted databases are collected from experienced professional artists. They are standard and most commonly used emotional databases. The complete range of emotions is available and results can be easily compared. These databases contain high-quality audio and they avoid recording problems such as microphone distance, codec-effect, noisy, and reverberant data. The SER accuracy is best achieved when the training and testing are performed on the same database. However, the accuracy significantly decreases when the training and testing are performed on different databases. Recently, various methods were proposed to deal with this problem. In Section 6, such techniques related to language-independent emotion recognition are discussed.

2. Evoked database: Evoked emotional speech databases are collected by generating an emotional situation. Here, a speaker converses with an anchor, who generates various contextual states to produce induced emotions in the speaker, which is reflected in the speaker's speech. This may not produce all categories of emotions. These databases are near to the natural databases and contain contextual information, however, the context is artificial. Tawari and Trivedi [36] studied the role of context in speech emotion recognition.

3. Natural database: The natural databases are naturally expressed and useful for real-world emotion modeling. These databases are collected by call center conversations, cockpit recordings, patient-doctor conversations, and public place recordings. Again, these may not contain all emotions. The problems associated with natural databases are copyright and privacy issues, uncontrolled noise, overlapping talks, and simultaneous occurrence of multiple emotions. Since the length of a naturally spoken utterance

Table 3
Summary of important databases.

Database	Type	# Recordings	# Speakers	Sampling rate (KHz)	Emotions
IEMOCAP [32]	Acted (Multi-modal)	10309	10	16	Anger, Happiness, Boredom, Fear, Neutral, Surprise, Excitation Frustration, and others.
AFEW [29]	Acted (Multi-modal)	1426	330	44.1	Anger, Disgust, Fear, Sadness, Happiness, Neutral, and Surprise.
RECOLA [31]	Natural (Multi-modal)	7 hours	35	44.1	Arousal degree (1-5), Valence degree (1-5).
CHEAVD [30]	Natural (Multi-modal)	2600	238	44.1	26 Non-prototypical emotions + 6 basic emotions and Sadness.
FAU-Aibo [39]	Natural	18216	51	16	Anger, Emphatic, Neutral, Joy, and Rest.
SUSAS [40]	Natural	16000	7	8	Low stress, Middle stress, High stress, and Neutral.

can not be predefined, deciding the unit-size (length) for emotion labeling is a challenge. The unit-size may be chunk, word, turn, or clips. Different databases use different unit-sizes for labeling. However, the effect of unit-size on emotion is not studied. Further, different annotators may assign different labels to a speech-unit, i.e., there can be a lack of consensus among humans regarding the emotion-label of a speech-unit. This problem is more prominent for natural databases. For example, as per the study carried out in [37], humans could perceive only 80% of the emotions correctly. A good thing about natural databases is their abundance, which is useful for deep-learning framework. However, such databases are imbalanced in terms of class and utterance lengths. To deal with these issues, recently some techniques were proposed as discussed in Section 5. The recordings of natural databases may suffer from problems such as noise, microphone distance, and channel mismatch; Section 6 discusses the techniques to deal with such problems.

The accuracy of a database also depends on the number and types of emotions present in the database. For example, the acoustic features of sad and boredom emotions are very similar, therefore, the classifiers get confused easily. Hence, the results of different databases cannot be compared because different databases contain different number and types of emotion. The emotional databases are labeled as per the two popular schemes: discrete categorical label (i.e., labeled as happy, anger, neutral, and sad) and continuous dimensional label. In the continuous dimension labeling approach, each emotion is defined as points in a 2-D or 3-D emotional space. The attributes are valence (negative-positive), activation (excited-calm), and dominance (strong-calm). For example, the anger emotion has a negative valence, high excitement, and strong dominance. Mencattini et al. [38] highlighted the usefulness of dimension labeling approach for applications such as diagnosis and surveillance. Recently, various databases have been constructed with continuous dimension labeling approach that will be useful for such applications.

IEMOCAP database [32] in spite of being acted is imbalanced and contains variable-length utterances to incorporate the naturalness. The texts of utterances are not the same in the IEMOCAP database, i.e., it is text-independent. Therefore, there is a big gap in the performance of the proposed methods on IEMOCAP database as compared to the other acted databases. The IEMOCAP database contains both scripted and spontaneous conversation of multiple speakers in dyadic sessions. In this way, it also contains contextual information while other databases lose contextual information. It is a multi-modal emotional database that contains audio, video, text, and gesture information to increase the emotion recognition accuracy, while other acted databases contain only speech. Due to the naturalness and more information, researchers are using IEMOCAP database increasingly to evaluate their results.

Recently, various emotion recognition contests, such as audio/visual emotion challenge (AVEC) and emotion recognition in the wild (EmotiW, focused on multi-modal emotion recognition results), have come up on natural or natural-like databases, such as AFEW, RECOLA, CHEAVD, and IEMOCAP [29–32]. These databases are more generalized because they are text-independent, are of varying length, and from speakers of varying ages. In such cases, the SER accuracy is less due to the inherent problem of natural databases. This survey can be useful in finding the cause of less accuracy in such cases and providing appropriate solutions. Table 3 summarizes important databases for SER in a natural environment.

3. Emotion modeling

Emotion recognition can be modeled both as a classification and regression problem. Emotion is a continuous activity, i.e., the level of any emotion is a continuous variable. For convenience, the psychologists have categorized the continuous variable into discrete emotions. There are seven basic discrete emotions: anger, happy, neutral, sad, frustration, disgust, and boredom. There are various applications [2,4] where the need is only to identify the discrete emotions. But some applications require the identification of the emotion level as a continuous variable, e.g., in disease recognition [6], and surveillance systems [9] in the identification of important calls at call-centers [5]. The identification of the emotion level as a continuous variable is done in terms of its three basic primitives [32]: (1) arousal (2) valence, and (3) dominance. Each primitive is a continuous variable that contains the value in a given range. The value of arousal captures the attentiveness; a low value indicates low attentiveness, while a high value points to high attentiveness. The value of valence captures the positivity of an emotion, a low value indicates low positivity, while a high value indicates a highly positive emotion. The value of dominance captures the power of emotion; a low value points to a low powered emotion, while a high value emotion points to a high powered emotion. The three primitives represent an emotion in a 3-D space and any discrete emotion can be interpreted in this continuous 3-D space. For example: high arousal, low valence and high power can be interpreted as anger emotion.

4. Features

The next step after the selection of a database is the extraction of features from the speech signal. Additionally, for natural databases, the speech signal needs preprocessing before feature-extraction. There are two reasons for this. First, while in general an emotional speech is labeled at the utterance level, the labeled emotion is not present in the entire utterance. Therefore, preprocessing is required to find the salient segments that are specific to

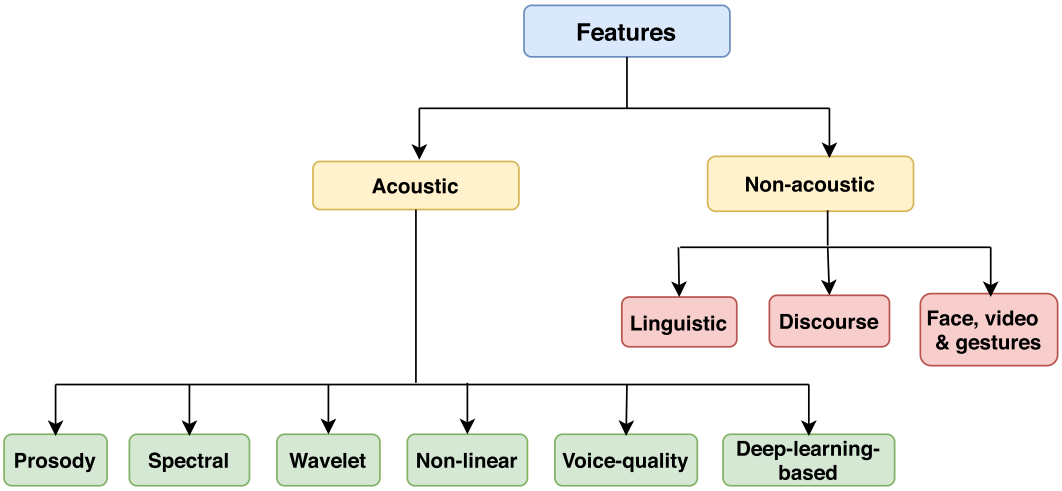

Fig. 3. Categories of emotion features.

the corresponding emotion. The second reason is the presence of background noise in the speech signal, which is removed during pre-processing.

The two factors do not affect the other types of databases because an emotion is consciously acted throughout the utterance by the actor and the recording happens in a controlled noise-free environment. Therefore, other types of databases need not require preprocessing.

The next step in speech emotion recognition is to find features that can discriminate different emotions properly. Here, a common practice for the task of emotion classification is the use of features popular with other applications of speech, such as speech recognition or speaker-identification. While such features work well for text/speaker/language dependent emotion classification, their performance accuracy dramatically decreases during text/speaker/language-independent emotion classification. This is due to the mismatch of the training and testing distributions due to different languages, texts, and speakers between the training and testing sets. Therefore, the features that are specifically designed for emotion recognition are discussed in Section 4.2. Further, the extracted features are often very high dimensional, resulting in over-fitting and requiring more time during training. This is solved by reducing dimensionality of extracted features. Next, the above-mentioned steps are discussed in detail.

4.1. Preprocessing

There are multiple approaches to segmenting speech signals: (i) with respect to the use of vocal cords as voiced and unvoiced, (ii) with respect to the unit of spoken text as sentence, word and syllable, and (iii) with respect to phonemes as vowel and non-vowel. Voiced segments are produced by the quasi-periodic vibration of the vocal cords, whereas unvoiced segments are produced either by turbulent airflow at the constriction or by the release or closure in the vocal tract [41]. In literature, mostly voiced regions are used for feature extraction [34]. Rao and Koolagudi [42] studied the effect of emotion at the sentence, word, and syllabic levels with respect to their position using prosodic features. The authors concluded that the prosodic features of the last word and syllable contain important and discriminative information for emotion discrimination. Koolagudi and Rao [28] showed the importance of transition region for emotion recognition. While agreeing with the importance of transition region, this work also claimed that pitch-synchronous spectral features are more logical than block-processing spectral features for emotion recognition. Deb and Dandapat [34] switched between vowel and non-vowel like regions to find the regions best suited for specific emotions. They concluded that anger, happy, neutral, and sad have high accuracy using vowel-like regions, while frustration and excitation have high accuracy for non-vowel like regions. From the above discussion, it is found that the identification of a segment plays an important role in the emotion recognition, in natural setting. Beyond the segmentation of speech signals as discussed above, work has been done to identify the salient segments specific to the corresponding emotion [43].

For removal of background noise – a common irritant in natural environment –, various speech enhancement techniques were used by Clavel et al. [9] and Stasiak and Rychlicki-Kicior [44]. This happens before feature extraction. The detailed description of these techniques is given in Section 6. It would appear that robust feature extraction can eliminate the need for speech enhancement techniques altogether [45,46].

4.2. Feature extraction

In this subsection, several acoustic and non-acoustic features are discussed for emotion recognition. The features that are extracted from speech signals are known as acoustic features, while the features extracted from non-speech elements are known as non-acoustic features. Further, the categorization of acoustic and non-acoustic features are depicted in Fig. 3. These features are not only used for controlled environment SER in a natural environment, but are also popular for SER in natural environment [23]. The acoustic features are explained next.

4.2.1. Acoustic features

Motivated by the accepted norm that the speech signal is non-stationary, it is first segmented into frames of size 20-30 ms and then features are extracted from these frames. These features are called low-level descriptors (LLD) features. The length of the utterance varies across databases. Due to the varying nature of the utterance length, the number of frames also vary for each utterance in the database. The utterance length is fixed by taking the statistical descriptors of each LLD feature across all frames in an utterance. The local level features are extracted frame-wise and the corresponding statistical descriptors (mean, median, max, min, skewness, etc.) of each feature are computed for the entire set of frames in an utterance. These statistical descriptors are also called global features. Global features, being lesser in number, have advantages over local features in terms of classification accuracy and time. In this way, the combination of local and global features [42,43,47]

are used for emotion recognition. In this paper, the acoustic features are divided into six categories, namely, prosodic, spectral, wavelet-based, voice-quality-based, nonlinear, and deep-learning-based (discussed in Section 4.2.1). Based on the origin, the non-acoustic features are categorized into linguistic, discourse, video, face, and gesture. This survey discusses, albeit briefly, the non-acoustic features and the corresponding circumstances in which these features can be useful (discussed in Section 4.2.2).

1. Prosody features: Prosody features are also called suprasegmental features because the frame duration of the speech analysis is large, typically 30-100 ms. Duration, intensity, the contour of the fundamental frequency (F0), and long term spectral features are mostly used as prosody features for emotion recognition. Busso et al. [48] used various statistics derived from F0 such as mean, maximum, minimum, range, standard deviation, skewness, kurtosis, and median for the emotion recognition task. The derivative of the F0 contour is computed to capture the dynamics of the intonation a long time. The intensity patterns of different emotions are measured in the same way. The different duration-based features are calculated on the basis of the relation of voiced, unvoiced, and silence segments.

The prosody features discriminate well between the low and high arousal emotions (e.g., between happy and sad), but they fail to discriminate among the emotions belonging to the same arousal (e.g., between angry and happy) or valence group (e.g., between sad and fear).

2. Spectral features: The spectral features are called segmental features because the frame duration of the speech analysis is about 10-30 ms. These features extract the energy content of different frequency bands of speech signal [49]. The important spectral features that were previously used for emotion classification are formant features [50] and cepstral features [51]. Cepstral features can be Mel frequency cepstral coefficients (MFCC) [52], linear predictive cepstral coefficients (LPCC) [53], and perceptual linear prediction (PLP) [54] coefficients. These features are calculated by the inverse spectral transform of the log-spectrum and capture the changes in the spectrum. The spectral features are also widely used for other speech applications. These features model the vocal-tract and are used for speech recognition, speaker identification, and emotion recognition [55]. However, they don't perform well in the case of speaker/text/language-independent emotion recognition. Recently, new spectral-based features were proposed [56,57] in the context of speech emotion recognition and they outperformed the SER accuracy of the existing MFCC, LPCC, and PLP features.

Rammohan and Dandapat [58] proposed a sinusoidal model-based analysis and extracted the amplitude, frequency, and phase features for recognizing different types of stress. Two models were developed, one based on vector-quantization classifier and the other on hidden Markov classifier to evaluate the chosen features. The best SER accuracy achieved 24.6% improvement over the LPCC features. Wang et al. [56] proposed novel Fourier parameter-based harmonic features. They first extracted frame-level harmonic features and then computed the statistics of these frame-level features along with their derivatives. The extracted features were normalized for each speaker using z-normalization [59]. The accuracy of SER was improved by 16.2%, 6.8%, and 16.6%, respectively on the German, CASIA, and EESDB databases when compared with the MFCC features and using the Fourier parameters. After combining the Fourier parameters with MFCC, the accuracy was improved up to 17.5%, 10%, and 10.5% on the German, CASIA, and EESDB databases, respectively.

Tao et al. [57] proposed spectral features based on the local Hu moments of Gabor-spectrogram. These features are computed by the following steps: (i) computation of log-energy of the spectrum, (ii) computation of Gabor-spectrogram by convolving the logarithmic energy spectrum with the Gabor wavelet, (iii) computation of Gabor local Hu (GSLHu) moments spectrogram, which is obtained through block Hu strategy and the subsequent application of discrete cosine transform to decorrelate the features, and (iv) finally, the use of PCA to eliminate the redundancy in the features. The proposed feature, called GSLHu-PCA, was compared with the MFCC, Hu-weighted spectral, and PLP features. The SER accuracy using GSLHu-PCA was improved by 7.47%, 4.97%, and 1.35% over the MFCC, PLP, and Hu-weighted spectral features, respectively, on the ABC database. Similar trends were observed for the EmoDB database. Luengo et al. [60] studied the effect of the prosody and spectral features. They claimed that the spectral features outperformed the prosody features. The combination of prosody, spectral, and voice quality features improved the result because these features capture different aspects of emotion. This work also combined the same features at different scales, namely segment and supra-segment levels, resulting in a better performance. Deng et al. [61] explored the phase features for SER. They proposed modified group delay and all-pole group delay features. The outer product of the trajectory matrix was used to fix the length of the feature vector, independent of the utterance length. Further improvement was obtained when the phase features were combined with the magnitude features.

Almost all the newly developed methods for extracting spectral features have the following in their techniques. First, they find the sub-band, which is mainly responsible for emotions. In this regard, it has been established that higher sub-bands of energy have a greater say in speech recognition than emotion recognition [62]. Second, they apply the statistical function upon the spectral envelopes. The Hu moments [57] is a statistical method that provides an excellent measure to discriminate emotion.

The spectral features capture the frequency content of different emotions. They give promising results for N-way emotion classification. The spectral features discriminate among the emotions having the same arousal and valence category. These features are highly correlated to the valence level in 2-D space [21]. The combination of the prosody and spectral features improves the result because the prosody features discriminate between the inter-arousal and inter-valence group, while the spectral features distinguish between the intra-arousal and intra-valence group emotions. Some works have also highlighted the usefulness of phase features for emotion perception and they have proven their complementary nature to spectral features.

Acoustic features are classified into time domain and frequency domain features. The time-domain features are computed directly from the signal waveform. Examples of time-domain features are frame energy [43] and zero-crossing rate [63]. However, a speech signal is better analyzed in the frequency domain. There are various ways to transform the time-domain signal into the frequency domain. Most of the frequency domain features are extracted using the Fourier transform. However, the Fourier transform only gives amplitude with frequency function, it does not provide the time instants at which these frequencies have occurred. This motivates the need for features that capture both spectral and temporal representation of speech signals simultaneously.

Wu et al. [49] proposed spectro-temporal features. These features are obtained from a long-term spectro-temporal representation based on the auditory inspiration. The representation is obtained using auditory and modulation filter-banks that capture both acoustic frequency and temporal modulation frequency components. Both the components are important for human speech perception, but the temporal information is missing from the conventional short-term features. As reported in their paper, the accuracy of SER was improved by 6.8% and 6.6% using the MSF (modulation spectral features) when compared with the MFCC and PLP features, respectively. The short-time Fourier transform (STFT)

overcomes the limitations of the traditional Fourier transform and captures both the temporal and spectral information. The STFT magnitude spectrogram, which is a function of time, is widely used in deep learning techniques for SER.

3. Wavlet-based features: Fourier transform follows the uncertainty principle because it cannot simultaneously capture both the time and frequency components. The window length has to be fixed a priori to compute the STFT of the signal. However, the fixed window-length adversely affects the representation because a fixed window does not capture the spectro-temporal representation for the entire speech appropriately. To address this, wavelet-transform is used in which the signal is decomposed into high and low-frequency components. Wavelet-transform provides better time resolution for the high-frequency components and better frequency resolution for the low-frequency components. An emotional speech varies rapidly with time than neutral speech. The details and fast changes in an emotional speech can not be analyzed properly with the fixed-length frame of STFT. Therefore, the wavelet transform (WT) is an alternative way to represent, decompose, and reconstruct the signals. It highlights the instantaneous changes in the spectral evolution. It is achieved by the discrete wavelet transform (DWT). Further, the DWT is extended by the wavelet packet transform (WPT) [64], wavelet perceptual packet (WPP) [65], and synchro-squeezing wavelet transform (SSWT) [66].

Recently, the wavelet transform has attracted the attention of the research community for SER. Huang et al. [67] proposed features based on wavelet packets (WP). They used a tree pruning algorithm to optimize the wavelet packet tree for the emotion classification task. The optimal sub-bands were selected and then the discriminative band WP power coefficient (dB-WPPC) features were derived. The accuracy of SER was improved by 5.1%, 8.16%, and 5.61% when compared with the MFCC, PLP, and Mel wavelt packet features, respectively. The wavelet packet tree was pruned using three criteria: Bhattacharyya distance (BhD), Fisher, and sub-band energy. Among these three criteria, the best result was reported using Fisher. Palo and Mohanty [68] extracted the MFCC and LPCC features based on Wavelet-analysis. The combined feature set was reduced with vector quantization approach and fed to the radial basis function network (RBFNN) classifier. The SER accuracy was improved by 23.49% and 22.32% using the proposed features when compared with the MFCC and LPCC features, respectively. Deb and Dandapat [35] proposed multi-scale amplitude features. They showed the superiority of these features over the traditional MFCC, Teager energy operator (TEO)-based, and breathiness features. The wavelet decomposition was used for multi-resolution analysis. After that, the signal was reconstructed into four sub-bands having bandwidths 1, 1, 2, and 4 kHz, respectively. For each sub-band, the sinusoidal model was applied. The noise part was removed using the threshold criteria of three descriptors namely the normalized bandwidth, normalized duration descriptor, and median pitch. Corresponding to the sub-bands having bandwidths 1, 1, 2, and 4 kHz, 4, 8, 16, 32 sinusoidal peaks were detected, respectively, and concatenated as the feature vector. The SER accuracy using the multi-scale amplitude features was improved by 5.9%, 9.3%, and 1.8% when compared with the breathiness, TEO, and MFCC features, respectively.

The wavelet-transform approximates non-stationary data using the basis function called wavelet. The wavelet needs to be selected a priori for linear decomposition, which should be orthogonal and complete. It can be drawn from a large dictionary of wavelets. Being a linear decomposition, the wavelet transform also suffers from the uncertainty problem. It is non-adaptive because the wavelets are selected a priori. Consequently, the wavelet decomposition is prone to spurious harmonics and misinterpretation of the data. In summary, the wavelet-based features are useful for emotion classification, however, one has to select a suitable wavelet for emotion recognition.

4. Voice-quality features: Voice quality features are also called sub-segmental level features because the segment duration of the speech analysis is less than 10 ms. An emotion is generated in the source signal as the constriction between the vocal cords varies while the air is expelled through the glottis. Kim et al. [69] studied different articulators for generating emotional speech. Their work may be useful in generating emotionally salient features for speech emotion recognition. Different phonations are generated by varying the glottal aperture. They are categorized as voiceless and voice phonations. The whisper phonation is voiceless while the modal, creaky, breathy, and falsetto phonations are voice phonations. There is a correlation between the different types of phonation and emotions [70]. For example, breathy voice has been associated with intimacy, harsh voice with anger, whisper voice with confidentiality, and creaky voice with boredom.

The voice-quality features [63,71] measure the attributes related to the vocal cords. These attributes may be the irregularity of the vocal cords, duration of opening and closing of the vocal cords, ratio of the duration of the opening and closing of the vocal cords, and relationship among the vocal excitations due to the vocal cords. Pribil and Pribilova [72] used shimmer and jitter features for emotion recognition. Shimmer measures the information about the temporal variation of amplitude, while jitter gives information about the temporal variation of F0.

Krothapalli and Koolagudi [73] derived features from the epoch locations. Epochs are discontinuous locations at which the vocal cord completely closes. The epoch-based features, namely instantaneous pitch, phase, and energy are used to recognize emotion. Authors combined the epoch-based and MFCC features to improve the accuracy of emotion recognition. Zao et al. [74] proposed the pH time-frequency vocal source feature and selected the region that supports the particular class by applying the acoustic mask to the amplitude modulation spectrogram. The proposed pH vector is a nonlinear feature where Hurst exponent is calculated using the wavelet-transform. The SER accuracy was improved by 6.8% and 17.7% using the proposed features when compared with the MFCC and TEO-based features, respectively. The auditory mask enhances the accuracy for both the conventional and proposed features.

Harmonic to noise ratio (HNR) [75] and normalized noise energy (NNE) [76] measure the relationship between the harmonics and noise present in the vocal tract. Glottal noise excitation (GNE) [77] ratio measures the impact of the turbulent noise during the vibration of the vocal cords. It differentiates among emotions based on the valence level rather than the activation level. The voice quality features may be useful for milder emotions because they discriminate among emotions having a common valence group. Therefore, these features can be combined with the prosody and spectral features to improve the emotion recognition accuracy. Mirsamadi et al. [43] combined voice quality features such as HNR and jitter with pitch (F0), voicing probability, energy, zero-crossing rate, Mel-filterbank features, MFCCs and formant locations/bandwidths for SER. We believe the voice quality features may be robust against the speaker, text, and language variations.

5. Non-linear features: During the production of an emotional speech, a non-linear pressure is exerted at the vocal cords; in this way the emotional speech can be considered to be produced by a non-linear system. This property of emotion cannot be captured using the classical features. In order to capture this property, the nonlinear dynamic (NLD) analysis was performed [74]. The NLD features capture the complexity of speech signals of different emotions. The popular NLD features are correlation dimension (CD), largest Lyapunov exponent (LLE), Hurst exponent (HE), Lempel-Ziv complexity, and entropy measures.

Tamulevičius et al. [78] proposed a feature set based on fractal dimensions (FD). They used the Katz, Castiglioni, Higuchi, and Hurst exponent-based FD features for emotion classification tasks having two to seven emotions. They showed the superiority of these features over the conventional features. Mao and Chen [79] used the correlation density (CD) and fractal dimension (FD) for emotion recognition. The importance of the CD and FD features was shown using the feature selection algorithm. The fractal dimension-based characteristics allow to reliably separate anger and happiness, which is not the case with the conventional features. The fear and boring emotions are still difficult to discriminate. Cairns and Hansen [80] used the Teager energy operator (TEO)-based features, which are also derived from the nonlinear assumptions for the stress condition.

The speech production process involves several physiological behaviors. Turbulent noise is produced by air passing through the vocal cords. Different paralinguistic behaviors are generated to characterize emotions due to the laryngeal tension. This operation produces a non-stationary nature in the emotional speech. This non-stationary nature of the emotional speech cannot be identified using conventional acoustic features. The wavelet transform addresses this problem, however, it decomposes the signal linearly, thereby requiring the nonlinear dynamics of the system to be captured using empirical mode decomposition (EMD). The EMD-based signal decomposition is nonlinear, non-stationary, and data-adaptive. It decomposes the speech signal into various modes represented through time domain. Each mode contains unique and meaningful information. Sharma et al. [62] combined the MFCC and EMD-based features for emotion recognition and reported improved performance when compared with the MFCC features.

6. Deep-learning-based features: Nowadays, deep learning algorithms are very popular. In deep learning algorithms, the LLD features are directly given to a convolutional neural network (CNN), long-short-term memory network (LSTM), or a combination of both. The deep learning algorithms learn in a hierarchical manner, learning higher-level abstractions of the low-level features. A CNN network learns spatial features, while the temporal relationship among features is learned by an LSTM network. The existing literature [81,82] has established that spectrogram gives better results than the LLD features because the LLD features are already decorrelated.

Liang et al. [83] manually extracted the color, direction, and brightness map of the spectrogram. These features were normalized by taking the average of small sub-matrix. Further, PCA was used to remove the redundancy from the feature set. The proposed feature set outperformed the traditional spectrogram-based features. There is another work by Ozseven [84] in which the spectrogram feature was extracted using the texture analysis. The texture analysis identifies the structural characteristics of an image. The texture is described as a spatial variation of the pixel intensity. The Gabor filter (GF), histogram of oriented gradients (HOG), gray level co-occurrence matrix (GLCM), and wavelet decomposition (WD) features are extracted from the spectrogram image. These features are complementary to the traditional acoustic features. Hence, these features significantly improve the result when combined with the acoustic features. However, these features can be extracted automatically using a deep neural network.

Chen et al. [85] used a CNN-LSTM model for emotion recognition. They showed the saliency of the delta and delta-delta log spectrograms for irrelevant factors such as speakers, speaking style, and environments. Sun et al. [86] utilized the advantages of both the shallow and deep features in a convolutional neural network. They claimed that the combination of the shallow and deep features produced a better result than the individual features. The deep learning-based features minimize the use of signal processing and their performance is comparable to the features using conventional speech processing. However, it is an open question whether these features can be replaced with the conventional acoustic features or not. For speaker and text-independent emotion recognition, the robustness of glottal spectrogram – when compared to a conventional spectrogram – was shown in [22]. Concluding the discussion so far, there is an opportunity for the researchers to generate a proper representation of speech signal for deep learning networks. Every utterance has a different number of frames. This creates a problem for deep learning networks because they require fixed-length inputs. The approaches that deal with this problem are discussed in Section 5.

7. Non-linguistic vocalization: It consists of non-verbal activities – such as laughter, breathing, irregularities, crying, and various breaks – that can be useful for emotion recognition [87]. These activities are also called speech disfluencies and can be detected using an automatic speech recognition engine. Yenigalla et al. [88] combined the spectrogram-based CNN features with the phoneme embedding features to deliver better performance. The phoneme embedding features are complementary to the spectrogram features because they capture various disfluencies present in different emotions.

4.2.2. Non-acoustic features

In the past, most of the emotion recognition studies focused on the single-modal recognition, such as the recognition of expression [89], speech [50], limbs [90], and physiological signals [91]. The absence of sufficient single-modal emotional data and the vulnerability to multiple external variables have led to low emotion recognition accuracies in such cases [92]. Consequently, in the affective computing field, the multi-modal information fusion for the data-driven emotion recognition has attracted quite a lot of attention [92–96]. Non-acoustic emotional cues, such as facial expressions or specific words, are helpful in many situations to understand the emotion of the target speaker [97]. The integration of the acoustic features with those based on linguistic, face, discourse, video, and gestures are used for emotion recognition. Due to a multitude of modality, such systems are called multi-modal emotion recognition systems. The non-acoustic features are discussed briefly next.

1. Linguistic features: Linguistic features for SER are those features that are extracted from the text-data associated with the speech of a person. These can be transcript of the speech or any other related text. They are unstructured, therefore, proper processing and analysis is required to find relevant linguistic features corresponding to a particular emotion. Strictly speaking, emotion analysis using text data is quite different than SER, and has a literature of its own [13,98,99]. Hence, in this section we just make a passing reference to some of the popular linguistic features for emotion recognition. They are bag-of-words (BOW), term-frequency (TF), term frequency-inverse document frequency (TF-IDF), self-referential information (SRI) and word-embedding (see [98,99] for further discussion). There are various works [98,100–103] in which the linguistic features were combined with the acoustic features.

2. Discourse features: In order to achieve a harmonious communication between computers and humans, it is essential to determine how a speaker is involved in conversations. Muszynski et al. [104] established the relationship between the perceived and induced emotions of the movie audiences using the discourse information. Majumder et al. [105] is another work in which the information of the past and future utterances were used to improve the SER accuracy using a recurrent neural network. Speech acts such as repetition, rephrase, and rejection of the utterance exhibit emotion-specific characteristics.

3. Video, face and gestures features: There are various works [93,106–109] in which these features were combined with the acoustic features. Rozgic et al. [94] combined the lexical, acous-

tic, and visual features, producing an improved performance on the IEMOCAP database. Tziraki et al. [95] used both speech and visual information for speech emotion recognition using a deep-learning framework. For video, face, and gestures, CNN-networks are more popular than other modalities. A CNN network automatically learns filters, and the output of the learned filters provides discriminative features.

Multi-modal emotion recognition has gained significant popularity in the last decade. The key challenges in relation to the multi-modal emotion recognition are: (1) how to combine multi-modal information appropriately and (2) how to deal with the information missing due to the privacy and other technical issues. To deal with the first challenge, Jiang et al. [92] discussed feature-level, model-level, and decision-level fusion methods to combine the multi-modal information. The second challenge is more complex and little work has been done so far. In one of the works, Aguilar et al. [110] efficiently combined acoustic and lexical modalities during the training phase while still providing a deployable acoustic model that does not require lexical inputs.

The multi-modal fusion may be useful in the natural environment because in the combined information the limitations of one modality are taken care of the features of the other modalities. For example, this survey discusses the issues of speech emotion recognition such as the independence of speaker/text/language and the uncontrolled environment (noise, microphone mismatch, etc.). The facial expressions can help in making up for the other modalities because they are universal and almost independent of speaker/text/language. However, some of these issues are also correlated with linguistic features. The linguistic features can be useful for solving recording issues, such as noises and microphone mismatch. In multi-cultural visual communication schemes, the recognition of facial emotions becomes vitally essential for emotional translation between cultures, which can be deemed analogous to voice translation [111]. Anger, happiness, and surprise emotions are video dominant, while sadness and fear emotions are audio dominant [112]. From the above discussion we can conclude that if the above two key challenges can be resolved, the multi-modal emotion recognition is a suitable choice for natural environments.

4.3. Feature selection

An SER system is an outcome of two stages: (i) a front-end processing unit that extracts the features from the data and (ii) a back-end unit that classifies the speech utterances into emotion categories. The dimensionality reduction techniques are an intermediate step between the front-end and back-end processing units, applied in order to reduce the high computational and storage complexities of the classifiers due to a high number of dimensions. These techniques also eliminate redundant features and preserve decorrelated features. There are two ways of dimensionality reduction: (i) feature selection and (ii) feature transformation. In the feature selection method, a subset of features is selected without changing the original feature values. The reduced set of features is selected such that they discriminate between different classes and do not vary within a class. The feature selection techniques can be categorized into filter methods [113], wrapper [114], and embedded [115]. The filter-based methods are simple, fast, independent of the learning methods, and inherently robust to over-fitting. The dimensionality reduction techniques are primarily based on the relevancy between the features and class, dependency between the selected and unselected features (redundancy), and the interaction between an individual feature and the selected feature subset (complementarity). Ideally, all the feature selection algorithms should satisfy the above three criteria. The first criterion is inherently built in the supervised algorithms because they use the label information. Most of the feature selection algorithms that are used for emotion recognition are supervised.

Pfister and Robinson [116] used the correlation-based feature selection (CFS) algorithm. It is a heuristic method that assumes that the selected feature subset should be highly correlated within the class and uncorrelated between classes. The CFS algorithm outperforms the wrapper-based feature selection algorithm, which evaluates each feature set by iteratively running the classifier. Gharavian et al. [117] used the fast correlation-based filter (FCBF) and analysis of variance (ANOVA) methods for the feature selection. FCBF is based on mutual information (MI). The author claimed that ANOVA outperforms the FCBF method for speech emotion recognition.

Chen et al. [118] discussed different feature selection techniques for a hierarchical classifier. In their approach, the feature subset was selected at each node of the hierarchy using Fisher ratio and principal component analysis (PCA). They pointed out that Fisher ratio is better than PCA. All their experiments were performed for six-class emotion recognition problem. Sun et al. [119] used the Fisher feature selection method in which the Fisher coefficients are obtained for each feature by calculating the mean and variance values of each feature. Features with higher coefficients values provide greater contribution to emotion recognition. Ozseven [120] suggested a new feature selection algorithm and compared this algorithm with the PCA, sequential forward selection (SFS), and FCBF algorithms in terms of the number of features and success rate. Their technique is based on the changes in emotions in the acoustic features.

Liu et al. [121] used a combination of correlation analysis and Fisher criteria for feature selection in speech emotion recognition. The assumption of this technique was to choose both the relevant and non-redundant features. The correlation analysis removes the redundant features and chooses the representative features. After this, Fisher coefficient is used for the dimensionality reduction. Gharavian et al. [122] used the FCBF feature selection method and genetic algorithm-optimized fuzzy ARTMAP neural network. Perez-Espinosa et al. [123] used the group-wise feature selection using the linear floating forward selection (LFFS) algorithm for 3-D speech emotion recognition. The authors claimed that the MFCC, LPC, and cochleogram groups are very prominent in estimating three emotion primitives. These three groups come under the spectral features. The energy group is only important for the arousal primitive. Ziang et al. [124] suggested a reordering of features with the weights fusion (RFWF) feature selection algorithm. It assigns weights to all the three criteria: relevance, redundancy, and complementarity. It reorders the features to find optimal feature subset. The proposed approach was shown better than the existing feature selection max-relevance min-redundancy (MRMR) and double input symmetrical relevance (DISR) methods, which do not give importance to all the three criteria. Demircan and Kahramanli [125] used the unsupervised fuzzy C-means clustering algorithm for dimensionality reduction. The reduced feature set showed a significant improvement over all considered features.

The feature transformation methods transform the original feature space into a lower dimensional space so that the features are decorrelated and the variance of the original feature set is preserved. The transformed features are not the same as the original features. You et al. [126] compared principal component analysis (PCA) and linear discriminant analysis (LDA) techniques, which are widely used for speech emotion recognition task. LDA is observed to be better than PCA for SER systems. Wang et al. [56] suggested a transformation method based on a sparse representation of scale-frequency map to reduce the emotional variance between training and testing data. However, both algorithms assume that features are linearly dependent. Ooi et al. [127] combined the bi-directional PCA and LDA for dimensionality reduction. The proposed frame-

work is better when compared to the standalone PCA and LDA, respectively. There are various dimensionality reduction algorithms that assume that variables are non-linearly dependent. These algorithms outperformed linear algorithms and are discussed next.

Vayrynen et al. [128] proposed a non-linear supervised dimensionality reduction method based on generalized regression neural networks using the prosodic and acoustic features. The assumption behind this technique is to choose those features that best discriminate between emotion classes in the target space with the specified embedding dimensions. The SER accuracy was improved by 2.9% when compared with PCA. Zheng et al. [129] proposed incomplete sparse least square regression (ISLSR) for emotion recognition. There are two advantages of their approach: (i) it uses both the labeled and unlabeled data, the unlabeled data is used for target domain adaptation, and (ii) it is capable of making feature selection where the choice of the features is made by imposing a $l_{2,1}$-norm penalty for the regression coefficient matrix on the objective function of the ISLSR. The ISLSR method achieved 8.38% better accuracy than Fisher+SVM on the reduced set of features. Yan et al. [130] suggested sparse partial least squares regression method (SPLSR) for the feature selection and dimensionality reduction of SER features. This approach outperformed the PCA, LDA, and PLSR methods. It is observed that only one group of data can be handled and analyzed by PCA and LDA, while canonical correlation analysis (CCA) and PSLR can handle two groups of data simultaneously. PSLR is unable to eliminate the redundant and insignificant features because the weight vector of PLSR is a linear combination of the entire speech features. The SPLSR method bettered accuracy by 14.93%, 4.23%, and 0.77% when compared to the PCA, LDA, and SPLSR methods, respectively, on the reduced set of features. Sahu et al. [131] proposed adversarial auto-encoder for speech emotion recognition. This network is capable of emotion recognition with very low dimensional space. The proposed method reduced the openSMILE (1582) features in two dimensions and improved accuracy by 7.71% and 13.26% when compared to PCA and LDA, respectively.

Concluding this section, it can be said that the non-linear feature selection algorithms work better than linear feature selection algorithms. Most of the feature selection algorithms use only emotion as the label information. However, one can use speaker, language, text, noise, microphone distance, etc., as well to find the subset of features that can work for natural environment [132]. Feature selection has been used to deal with the issue of the microphone distance in SER in natural environment. Another point is that any particular feature subset cannot be accepted universally to classify all emotions. Therefore, a hierarchical classifier is better, where at each node a different feature subset is selected for a particular emotion group. For example, the pitch related features are better for classifying happy and sad emotions, while the spectral features are better for classifying the happy and anger emotions. The unsupervised feature selection algorithms may be useful when the class information is unavailable.

5. Model

The next step is to create a model to classify the utterances. Many classifiers have been proposed in literature in general, and most of them have been used for speech emotion recognition. While the existing surveys have discussed them in terms of accuracy, this survey also discusses the different scenarios of SER in a natural environment under which certain classifiers can be better suited. Recently, the deep learning-based classifiers have shown commendable performance as compared to the conventional classifiers, while requiring much less signal processing. The classifiers based on deep learning are still in an early stage and need to be explored for their potential utilization in a natural environment. In this survey, the deep learning issues and related techniques are discussed in detail. While discussing the models we have emphasized on the elements helpful for SER in a natural environment, however, we have discussed models used for SER in both natural and non-natural setting, because the models used in SER in a non-natural settings also have application potential for SER in a natural setting.

5.1. Conventional models

A variety of machine learning algorithms, such as SVM, HMM, Gaussian mixture model (GMM), K-nearest neighbor (KNN), artificial neural network (ANN), decision tree, and fuzzy classifiers, have been used for speech emotion recognition. Each algorithm has its own set of advantages and disadvantages. There is a need to select the appropriate classifier for speech emotion recognition. The statistical classifiers (GMM and HMM [64,72,74,103,133,134]) are very popular for the speech-based applications. GMM is a simple classifier, it captures the probability distribution of each class using a mixture of Gaussian components. It is like continuous HMM with only one state. An advantage of HMM is that it is quite capable of modeling the temporal dynamics of emotions [135–138]. Recently, DNN-HMM models [137,138] gained popularity in the context of the speech-based applications. In a DNN-HMM model, the posterior probability of each state is calculated using a DNN instead of GMMs because GMMs are vulnerable to nonlinear data. This is useful because each emotion contains unique temporal behavior along with nonlinear data. This temporal behavior is related to the contextual information. Acted databases do not capture contextual information, while it is prominent in natural databases. Therefore, DNN-HMM models are specially suitable for SER in natural databases.

SVM is a widely used generalized discriminant classifier [33,44, 49,139–152]. The assumption behind SVM classifier is to reduce the training error as well as the empirical risk (testing) error. It uses kernel functions to map original feature space into a higher dimensional space so that it could be classified by linear SVM. The selection of kernel function is a crucial task in SVM. SVM is still popular for speech emotion recognition due to its generalization property and lesser vulnerability to outliers and high dimensional features. SVM classifiers also work well without feature selection algorithms. Other classifiers, such as decision tree, KNN and fuzzy classifiers [153–156], were also used for speech emotion recognition. These classifiers are sensitive to the high dimensionality of features – a common occurrence in SER, e.g., 1582 features in [157].

There are various works [33,158] where multiple classifiers are combined to gain accuracy. Despite the increase in the computational complexity for multi-classifier systems, they are used in SER for their better performance. Wu and Liang [158] combined three different classifiers namely GMMs, SVMs, and ANN, using a meta decision tree. They used the acoustic features, prosodic information, and semantic labels for recognizing the emotion. The experimental result revealed that meta decision tree (MDT) combining each individual classifier bettered accuracy by 7.39%, 1.84%, and 8.13% when compared to the GMM, SVM and multi-layer perceptron classifiers, respectively. Classifiers can be combined using two approaches: hierarchical and parallel. In the hierarchical approach [33], the classifiers are arranged in a tree-like structure. For example, at the root node, an emotion is classified into a broader category (e.g., high-arousal/low-arousal emotion). As one goes deeper in the tree, the classification becomes finer and the number of candidate classes becomes smaller. At the leaf node, a decision is taken for a single class only, which is the final classification output. This approach is motivated by human psychological behavior.

An emotion can be represented in a 3-D plane by three attributes, namely arousal (attentiveness), valence (positivity), and dominance (power). Using this representation, a hierarchical classifier is arranged as a 2-stage or 3-stage classifier. At each stage, i.e., at each node of the hierarchy tree, same or different classification algorithms can be applied. In the parallel approach all classifiers are independent. The final decision is made by combining the output of all the classifiers based on fusion techniques. The hierarchical classifier often gives better results due to a small number of classes at each node. Most of the works [33–35] using hierarchical classifiers have chosen the same set of features at each node. However, better results may be achieved if a different set of discriminative features are chosen at each node.

The multi-classifier system is also called ensemble classifiers (EC). Recently, SER systems based on ensemble classifiers [159–163] have produced promising results. This is because the ensemble classification reduces variance within the data and prevents over-fitting. This can be helpful for SER in a natural environment because, in general, natural emotional speech has large variance due to speakers, text, language, etc. However, ensemble classifiers require decorrelated features and parameters, which is a disadvantage.

5.2. DNN-based model

Artificial neural networks (ANN) are very popular in machine learning algorithms. They mimic the behavior of the human brain. There are three types of layers in a neural network: input, hidden, and output layers. There is only one input and one output layer, while the number of hidden layers may vary. If the number of hidden layers is more than one in a neural network, it is called a deep neural network (DNN). The more are the hidden layers, the more abstract representation of a pattern is learned by the neural network. Neural network algorithms effectively model the non-linear mappings. These algorithms require a large number of samples due to a large number of parameters in the network. The decision on the number of nodes in a layer, number of hidden layers in a network, optimizer, loss function, dropout rate, and learning rate is crucial. The broad categories of ANN algorithms used for SER task are feed-forward neural network (FFNN), convolutional neural network (CNN), recurrent neural network (RNN), and auto-encoder neural network (AEN).

Nowadays, most of the systems are end-to-end, where the selections of the important segments and discriminative features are automatically performed through the learning of the neural network [16]. Han et al. [164] used DNN model to extract features from the speech segment. Further, the utterance level features (global features) were constructed using statistics of the posterior probability. These utterance-level features are fed to the extreme learning machine (ELM) classifier. ELM [165] is a single hidden layer feed-forward neural-network (SLFN), where the number of hidden nodes are chosen arbitrarily and only the output layer weights are trained. It does not require the input layer weights and the first hidden layer biases to be tuned. It is fast and has good generalization capabilities, because of which it is preferred over the simple feed-forward neural network in SER [166]. The ELM network has only one hidden layer and is called as a shallow network.

There are other types of neural networks where more than one hidden layer can be used; they are called deep networks. Such types of networks learn different aspects of data such as CNN (Spatial), LSTM (Temporal), and AEN (Generative features). Recently, CNN is widely used for SER. There are various input representations used for CNN. Badshah et al. [167] used the MFCC and Mel-filter bank representation for CNN. These feature representations have smaller dimensions, so the model does not overfit with low resource data. Recently, the spectrogram is widely used as input for CNN. A spectrogram is a time-frequency representation of the signal. The time-varying short-time Fourier transform (STFT) magnitude spectrum is displayed in the form of an image, popularly known as a spectrogram. Mirsamadi et al. [43] highlighted the importance of the raw spectrogram over the hand-crafted features. The raw spectrogram has a higher accuracy than the hand-crafted features because the hand-crafted features are already de-correlated. In a spectrogram, the Fourier transformation of a short-time windowed signal is taken with overlapping. The length of the short-time window is 25 ms with overlapping of 10 ms. There are various existing works [82,167–169] in which spectrogram was used as input to the CNN.

Mao et al. [82] proposed salient features and a novel objective function using spectrogram and CNN, which are robust for speaker and language-independent emotion recognition. These features are also robust to environmental distortion. The SER accuracy was improved by 22.3% and 46.9% using the proposed approach when compared with the TEO and raw audio features, respectively. Zheng et al. [170] also used spectrogram with deep convolution neural networks (DCNNs). PCA was used to reduce the dimensionality and to suppress the interferences of log-spectrogram. The SER accuracy was improved by 2.41% when compared with the SVM+hand-crafted acoustic features. Lee and Tashev [166] used an LSTM-ELM network with hand-crafted features, where the LSTM network utilized the temporal dynamics and uncertainty of emotion in an utterance. The weighted and unweighted accuracy of LSTM-ELM framework was improved by 12% and 5%, respectively, when compared with the DNN-ELM framework. Zhao et al. [171] combined CNN and LSTM network, called CNN-LSTM, where the LSTM network learns the temporal sequence of features learned by CNN. Authors concluded that the spectrogram representations have a 6.62% higher accuracy than the raw audio signal. This work also highlighted the accuracy gap for speaker-independent emotion recognition. The CNN uses a filter to extract the local features, and these filters are learned during training through data.

Every part of the input representation is not related to the corresponding emotion. To deal with this problem, attentive CNN [176] was proposed that assigns respective weights to feature maps extracted from different parts of the input. The authors used the MFCC, Mel-filter bank, eGeMAPS, and prosody features. They found the performance of the Mel-filter bank representation the best. Fayek et al. [169] detected and removed the silence frames during training and testing. They segmented an utterance into frames, labeling the frames with the corresponding emotion of the utterance or with silence for silent regions. Frames were concatenated with the previous left frames of 2.6-sec length. Finally, the utterance was classified by averaging the maximum posterior probability and ignoring the silence segments. Li et al. [177] investigated the optimal size of filters from the accuracy perspective, finding the optimal sizes to be 10×2 and 2×8. The output of these two filters was concatenated. They used global average pooling and soft attention pooling after the last convolution layer. Simple concatenation results in over-parameterization and creates a problem for low resource databases. This problem is dealt with the use of the global average pooling method, which efficiently projects the feature maps into one-dimensional feature vectors with equal weights. The attention pooling method can be seen as an advanced version of global pooling that inherits the property of emotional utterance by assigning weights according to the salient segments.

Huang et al. [180] showed the importance of unsupervised learning in SER through the sparse auto-encoder and sparse restricted Boltzmann machines. Their experimental findings indicate that large patch size of spectrogram and hidden nodes contribute to higher SER accuracy. Sahu et al. [131] proposed adversarial auto-

Table 4
Pros and cons of different classifiers for speech emotion recognition.

Classifiers	Pros	Cons
Conventional classifiers		
SVM [33,44,49,139–152]	• It works well for small databases and high dimension features, which are common in speech emotion recognition. • It is a generalized discriminant classifier because it cares about testing data.	• The choice of kernel and its parameters and regularization of parameters to avoid over-fitting are crucial.
KNN [153–156]	• It is inherently able to model non-linear relation of emotion features.	• The choice of K and distance measures are crucial. • It is sensitive to outliers so feature selection must be used before KNN.
EC [159–163]	• The variance is large in emotional speech features due to the variance of speakers, texts, and languages, etc. • It reduces variance and prevents over-fitting.	• It does not work when emotional features are correlated. • The hyper-parameters such as depth of Xg-boost algorithm is crucial.
GMM [103,133,134] [64,72,74]	• It is generative and learns the latent distribution of a particular emotion. Hence, it can be combined with discriminative classifiers such as SVM.	• It does not work when the dimensions of emotional features are too many. • It is unable to model non-linear relationship among emotional features. • The number of Gaussian components is crucial to model the distribution of the emotion.
LDA [3,48,172]	• It reduces the dimensionality of input data, thereby reducing the computational burden.	• It is unable to model the non-linear relationship of emotional features. • It requires a large number of utterances.
HMM [135–138]	• It learns contextual information with less amount of data. • It is useful for a natural emotional database where context is available.	• The number of hidden states is crucial for varied length utterances.
Deep-learning based classifiers		
ANN [26,164,173,174]	• It effectively models the non-linear relation of emotional features. • It is suitable for real-time applications due to less prediction time.	• Require large resources and time to build emotion recognition model. • The hyper-parameters such as the number of hidden nodes and layers are crucial. • Requires attention to the important features and segments. • Requires a large amount of data.
CNN [82,88,167–171,175–178]	• Minimal signal-processing, automatic learning of discriminative and global features.	
LSTM [43,166,168,171]	• Learns long contextual information of natural emotions. • Deals with variable-length input utterances.	
AEN [27,148,179]	• Used to deal with mismatched environments in emotion recognition. • Learns generative features in low dimensional space, able to learn non-linear relationship of emotional features.	

encoder for SER. This network is used to reduce the dimension of the features so that it discriminates among different emotions. Table 4 summarizes the pros and cons of each classifier; the classifiers are categorized into two broad categories: conventional and deep-learning-based classifiers.

From the above discussion, it can be concluded that deep learning networks automatically learn the different aspects of input representation. Though the spectrogram is a better representation than all the LLD features and raw audio, still there is a need to explore other input representations for SER and the combination of spectrogram representation with the other LLD features. The combination of the CNN-LSTM network gives better results because it captures both spatial and temporal relationship of the input representation. The ELM network has good generalization capability, which is suitable for SER in a natural environment. However, the shallow nature of ELM restricts its use for SER. The combination of ELM with CNN and LSTM may be useful because of their individual strengths. The auto-encoder networks are useful to reduce the dimension of features. They learn generative features that may be combined with the derived features from the CNN-LSTM network. The size of filters, number of filters, number of hidden-layers, and number of nodes in the hidden layers vary across research papers. Deep-learning methods require a large number of samples to build an emotion recognition system. Such type of models overfit due to a large number of parameters and a limited amount of annotated samples, which is common with most of the databases. The attention mechanism is expected to find important segments and features for SER, however, any considerable improvement in accuracy is not observed [43,176]. There is a need to explore the deep-learning mechanisms to find salient segments and features. The labeled data should be large in quantity and balanced to implement deep-learning-based algorithms. However, most of the databases are too small, which is not suitable for deep-learning algorithms. Natural databases are imbalanced in terms of class and utterance lengths. Various techniques have been proposed to deal

with these issues, which are discussed in detail in the following section.

5.2.1. Deep-learning-based issues related to natural environment

1. Lack of data: The lack of a large natural database is one of the main obstacles in transferring results on SER in controlled environments into real-life applications. [181]. It is costly and time-consuming to build emotional databases, which limits the size of the existing corpora. In such a scenario, models developed based on deep-learning algorithms overfit due to a large number of parameters. A large amount of data can overcome this problem, however, most of the databases have a limited amount of annotated samples. To deal with these issues, either the number of samples is boosted by oversampling or unlabeled data is used to implement transfer learning. It is feasible to combine crowd-sourcing and machine learning algorithms to obtain a balanced emotional database with a cost-effective labeling procedure.

Weisskirchen et al. [182] used a different upper limit of high frequency to generate more spectrograms of the same utterance. Oversampling can also be achieved by perturbation of the vocal tract length (VTLP) [183] of the samples of the lesser represented classes in a database. This technique generates new samples by re-scaling the original spectrogram along the frequency axis. The scaling factor typically lies between 0.9 and 1.1. This process is an instance of data augmentation. The accuracy was improved by 4% when oversampling was used [183]. There are various databases available for SER. Similar databases are combined to increase the amount of data. Siegert et al. [184] used a PCA-based similarity measure to combine multiple databases. In their experiments, they found that the combination of the six databases: ABC [185], emoDB [186], eNTERFACE [187], SAL [188], SmartKom [189], and SUSAS [40] performed very well due to significant similarity among them.

The other approach is transfer learning, where the central concept is to transfer the latent distribution characteristics of the unlabeled data to labeled data. Once this is done, a deep learning model can be built with even less amount of labeled data. The popular algorithms that are trained with unlabeled data are maximum a posteriori probability (MAP) and autoencoder neural networks. The transfer learning approaches have been successfully used in SER [190,191].

2. Data imbalance: The data imbalance is an inherent issue of natural databases. The imbalance exists with respect to class, utterance length, etc. Tang et al. [192] highlighted the class imbalance problem and proposed an oversampling procedure. In this procedure, sampling without replacement is used for neutral emotion and sampling with replacement is used for other emotions. This strategy is used because the quantity of neutral emotion is more in nature. Shih et al. [193] proposed a robust neural network model where skewed weights were assigned to the training samples during the weight updation. Fewer weights were given to the majority classes, whereas more weights were assigned to the minority classes. The weights were assigned in inverse proportion to the number of samples in a class.

Bang et al. used SMOTE [194] to balance the minority classes. First, SMOTE finds the K closest sample neighbors of a minor class. It then computes the difference between the current sample and these K neighbors. This difference is multiplied by a random number between zero and one. The resulting product is added to both the original samples and the training data. This increases the number of samples of minority classes. The above process is repeated until the number of samples for all classes are balanced. SMOTE can also suffer from the cold start problem when there are a very few numbers of samples for a particular class.

3. Utterance length: With respect to the utterance length, there are two factors conducive for emotion recognition: (i) near-constant length of the utterances and (ii) absence of long utterances. The long utterances are tricky because, in the short or average length utterances, the same emotion is contained throughout, whereas, in long utterances, emotions may be present in only parts of an utterance. Therefore, there is a need to identify the emotional regions in the utterances. The training and testing should be done on emotional regions only.

Chang et al. [195] showed the effect of duration on emotion recognition. They partitioned the utterances in the database into different durations (short: 0.50–1.00 s; medium: 1.50–2.00 s; and long: 2.50–3.00 s). They found that the short duration utterances are more discriminative in different emotions than the medium and long ones. For conventional approaches, the processing of speech signals is performed frame-wise. After that, the global features are extracted from the local features by a statistical function. This process produces fixed-length feature vectors.

In deep-learning, the global features are learned using frames only. Due to the varying length of utterances, this creates unequal length inputs for deep-learning algorithms. This problem is resolved by either padding the smaller utterances with zeroes till the desired length or segmenting the utterances into fixed-length segments. At the time of testing, the posterior probabilities of the segments corresponding to an utterance are averaged. The utterance is assigned the class whose averaged posterior probability is maximum.

Zhang et al. [196] proposed discriminant temporal pyramid matching (DTPM) to find the utterance level features. First, an utterance was segmented into fixed-length segments, then the DTPM was used to extract the fixed-length feature vector for each utterance. The Mel spectrogram with their delta features was fed as input to the CNN classifiers. The DTPM outperformed other pooling strategies at the segment level and the padding strategy at the utterance-level. Concluding this section, the segmentation approach appears to be better than the padding strategy because during segmentation, padding happens only for the last segment.

6. General issues related to natural environment

Emotion recognition systems can be developed using various modalities. Each modality has issues in the natural environment. Multi-modal emotion recognition system resolves these issues by incorporating features from multiple modalities, where the limitations of the features of one modality are taken care of the features of other modalities. However, multi-modal systems do not work when the information of the other modality is absent. In such a scenario, one has to depend only on the modality for which information is available. In this survey, it is assumed that only speech modality is present.

6.1. Speaker-independent emotion recognition

The performance of SER systems degrades due to the use of different speakers during the training and testing phases. SER systems typically perform well when the same set of speakers are used for both the training and testing phases. However, in natural settings, most SER systems are likely to encounter a new test speaker on which the system has not been trained before. In such cases, i.e., training and testing with different speakers, the performance degrades dramatically. This is because of the anatomical and morphological variation in the vocal-tract geometry of different speakers [197]. The speaking rate also varies with the age group of the speakers. To discuss the approaches proposed to deal with this challenge, we have categorized the existing methods in the literature into four categories: (i) speaker-independent features, (ii) speaker normalization, (iii) speaker adaptation, and (iv)

Table 5
Pros and cons of different approaches used for speaker-independent emotion recognition.

Speaker independent emotion recognition				
Approach	Method	References	Pros	Cons
Speaker-independent Features	RSS features	Kim et al. [18]	• No Speaker information is required. • It can be applied to any classifier.	• Need for signal processing expertise.
	Bi-spectral features	CK et al. [19]		
Speaker-normalization	z-normalization	Sun and Wen [20]	• Simple and easy to implement. • No speaker information is required.	• Need to classify neutral utterances
	VLTN	Etienne et al. [183]		
Speaker-adaptation	Maximum a posteriori (MAP)	Yang and Hung [198]	• Requires little adaptation data.	• Supervised. • Specific to GMM and HMM classifiers. • More time is required. • Both emotion and speaker labels are required.
	Maximum likelihood linear regression (MLLR)	Kim et al. [18]	• Unsupervised and significant improvement in accuracy.	• Requires more adaptation data. • Speaker information is required. • Cold start problem. • More adaptation time.
Others	Ensembling	Schuller et al. [199]	• No speaker information is required.	• Complex and requires more time.
	Covariance-shift	Hassan et al. [200]		
	Discriminative training	Kockmann et al. [201]		

others. These approaches along with their pros and cons are summarized in Table 5.

1. Speaker-independent features: The goal here is to find speaker-independent features that do not vary with speakers. In this direction, Kim et al. [18] proposed RSS (ratio of the spectral flatness to the spectral center) features. The authors claimed that the RSS features are less variant with respect to the speaker and text when compared to the other existing features, i.e., the MFCC, pitch, and energy features. A hierarchical classification method was used in which at the first level the high arousal emotions were put in one group and low arousal emotions in another. At the second level, bifurcation was done between anger/joy and sad/neutral emotions. At the first level, the pitch, energy, and MFCC features were used, while the RSS features were used at the second level. The SER accuracy was improved by 8% for sad/neutral and 13.5% for anger/joy when compared to the MFCC features. Similar trends were observed for the LPCC and PLP features.

Yogesh et al. [19] proposed weighted bi-spectrum based features for stress and emotion classification from natural speech. Unlike the power spectrum, which is a function of one frequency variable, the bi-spectrum is a function of two frequencies. The 28 bi-spectral features were extracted from both the original speech and glottal signal and then concatenated. The emotion recognition rate reduced due to the mismatch of the signal to noise ratio (SNR) between the training and testing data. This is due to less inter-class variance and high intra-class variances among the features. A mean shift clustering algorithm was used to compute the weighted features by multiplying each feature with the ratio of the mean of the feature to its cluster center. The proposed features were robust against speaker, gender, and text-independent emotion recognition. The SER rate was improved by 36.49%, 36.49%, 37.33%, 32.55%, and 44.18% using weighted bi-spectral features for GRNN, PNN, KNN, and ELM models, respectively, when compared to the raw bi-spectral features. These improvements were achieved on speaker-independent emotion recognition. The merit of this approach is that the speaker information is not required for speaker normalization. However, this approach requires expertise in signal processing.

2. Speaker normalization: This approach is simple and easy to implement. There are a few research works [20,202] that have used speaker normalization for developing SER systems. Sun et al. [20] proposed a speaker normalization technique that consists of two steps. In the first step, the features are normalized using z-normalization by the mean and standard deviation of all the training utterances. This removes the effect of the varied range of features in the source data. In the second step, the features of the testing utterances of a particular speaker are shifted by their mean. The variance parameter is not used because the estimation of variance requires more data, which often cannot be guaranteed at the time of testing. The second step normalizes the differences among the speakers. The above method does not require the speaker identities for speaker-normalization. Recently, testing utterances have also been normalized using the statistics of the training data [168]. This is because the amount of testing data is often not sufficient to calculate the required statistics.

In general, only neutral data is used to compute the statistics of a particular speaker [168]. This is because the neutral data is abundantly present. Further, if all emotions are considered, the corresponding statistic measure will normalize the differences among the emotions, thereby reducing the discriminative ability of the features. In [59], z-normalization and range normalization were applied. The accuracy improved by 8.2% and 6% for z-normalization and range-normalization, respectively, when compared to without normalization. Busso et al. [203] proposed a method that solves the requirement of the neutral data of the testing speakers. The proposed iterative feature normalization framework (IFN) automatically decides whether the testing data is emotional or neutral. Normalization is then performed based on the statistics of the neutral data. This approach assumes that the speaker information is already known. Normalization is performed using the statistics of the neutral samples. This approach compensates for the acoustic differences among speakers, but it does not affect the acoustic differences among emotional classes.

Vocal-tract length normalization (VTLN) [183] is another method for speaker normalization, where the inter-speaker variability is reduced. The difference in the length of the vocal tract of individuals can be modeled by rescaling the peaks of important formants along the frequency axis with a factor α of the estimated range values (0.9, 1.1). To address this variability in the estimated values, first the value is estimated for each speaker. The estimated value is then used to normalize the speaker's spectrograms accordingly.

Table 6
Pros and Cons of different approaches used for text-independent emotion recognition.

Text-independent emotion recognition				
Approach	Method	Reference	Pros	Cons
Text-independent features	RSS	Kim et al. [18]	• Significant improvement in accuracy. • Applied for any classifiers.	• Not easy to extract such features. • Signal Processing expertise is required.
	Adaptive-TEO and Pitch	Wu et al. [21]		
	Glottal spectrogram	Ghosh et al. [22]		
Transformation	Whitening-transform	Mariooryad and Busso [23]	• Applied for any features. • Signal processing expertise is not required. • Less computation time. • Simple and easy to implement.	• No significant improvement in accuracy.
	Functional PCA	Arias et al. [206]		
Ensembling	GMM-DNN	Shahin et al. [24]	• Applied for any features. • Signal processing expertise is not required.	• Classifier specific. • More complex. • More time.

3. Speaker-adaptation: The third category of approach consists of speaker-adaptation techniques like maximum a posteriori (MAP) [198] and maximum likelihood linear regression (MLLR) [204]. In this approach, the model developed using training data is re-estimated using the new testing speakers. If some amount of labeled data is required for re-estimating the model, it is called supervised adaptation, else unsupervised adaptation. MAP is a supervised adaptation technique. In the case of SER, due to the almost certain unavailability of labeled data for testing speakers, unsupervised speaker-adaptation technique MLLR is more suitable than MAP. However, the identity of the speaker is required in both cases. MLLR has been used for personalized emotion recognition [204]. Fahad et al. [138] used a feature-based MLLR approach for speaker-adaptation and the accuracy was improved by 7.13% when speaker-adaptation was not used.

4. Other methods: The other methods that do not fall under the above three categories are described here. Schuller et al. [199] stacked the SVM, naive bayes, C4.5, and KNN classifiers. During experiments, they found this approach to be robust against speaker-independent emotion recognition. Such ensemble methods are robust because each model captures different aspects of emotions. Hassan et al. proposed [200] a method to minimize the covariance shift between the training and testing set of speakers. Weights were introduced in the SVM formulation to reduce the covariant shift. A high weightage is given to those training samples that match with the testing samples, while a low weightage is given to those samples whose distribution does not match. The accuracy was improved by 1.8% and 3.3% using the covariance shift method for two class and five class classification problems, respectively, when compared to the combination of cepstral mean normalization and vocal tract length normalization. Kockmann et al. In [201], a technique was proposed for speaker and language-independent emotion recognition. They used GMM in which discriminative training is performed using maximum-mutual-information (MMI) criterion and inter-session variability (ISV) compensation. The goal of all the above methods is to minimize the difference in distributions of the training and testing set of speakers.

6.2. Text-independent emotion recognition

The expression of an emotion is highly associated with the spoken text. It is impossible to include all possible text that might be spoken by the speakers during training. This causes text mismatch between training and testing data. In the early stages of research, most of the emotion recognition models were evaluated on the Berlin [186] and IITKGP-SHESC [205] databases. These databases contain only ten utterances that are repeated in both training and testing sets. Researchers achieved very good performance with their proposed approaches to these datasets. In comparison, several natural-like databases, such as IEMOCAP [32], contain utterances that are independent of each other and different utterances are present in the training and testing phases. It is not surprising that the achieved accuracies for the IEMOCAP database are less as compared to the other databases (e.g., Berlin, IITKGP-SHESC). The reason for this is the mismatch in the distribution of the spoken text in the training and testing datasets. The techniques that deal with this issue can be categorized into three high-level approaches: (i) Text-independent features, (ii) transformation, and (iii) ensembling. These approaches along with their pros and cons are discussed in Table 6.

1. Text-independent features: The goal here is to find features that do not vary with text. Wu et al. [21] proposed text-independent pitch and adaptive-TEO features. For computing these features, the formants frequencies were grouped into six clusters and pitch features were extracted from those utterances that belong to the same cluster. The TEO-based features were adapted using the sub-band information. The adaptive TEO-based features were robust against text-independent SER with the lowest variance value of 0.034 when compared with the traditional pitch feature (F0) and formant features (F1 to F5). In another work, Ghosh et al. [22] proposed glottal spectrogram. The glottal signal was extracted from the original speech, which was subsequently used for constructing the spectrogram. The glottal spectrogram showed robust performance for the text and speaker-independent emotion recognition. The weighted and unweighted accuracy were improved by 1.77% and 2.11%, respectively when compared with traditional spectrogram. This approach significantly improved accuracy and works well with various classifiers. However, a signal processing expert is required to find the text-independent features. The goal of the above-discussed methods is to choose such features that minimize the difference in the distributions of the training and testing data with respect to textual content.

2. Transformation-based methods: The goal of this approach is to reduce the variance between training and testing data [23,206] using feature transformation approaches. Mariooryad and Busso [23] used various acoustic features highlighting their dependencies on the lexical contents, emotions, and speakers. They further normalized the lexical and speaker factors for SER. They used whitening transformation for the speaker and text normalization, after which a reference phoneme was chosen. The statistics of reference phoneme was imposed on all phonemes in the database. The normalization approach reduced the variability against text and the relative performance was improved by 4.1%. This approach is simple and takes less time, but any significant improvement in

accuracy is not observed. Arias et al. [206] suggested a new technique for detecting the emotional modulation in the $F0$ contours using neutral reference models with the PCA-basis function. This approach is robust against text-independent emotion recognition and accuracy was improved by 6.2% when evaluated using without PCA-basis function on the same features.

3. Ensembling: The last approach is ensembling, where a combination of classifiers is utilized to extract the emotional attributes. Shahin et al. [24] proposed a text and speaker-independent emotion recognition approach using a hybrid cascaded Gaussian mixture model and deep neural network (GMM-DNN). In GMM-DNN, a separate GMM model was developed for each emotion using the training data. Then, the training data was passed through each GMM model and the log-likelihood distance between the training features and the GMM tag was calculated. The log-likelihood distance was used as a feature to train the DNN. The accuracy was improved by 3.7% and 14.2% respectively when compared with SVM and MLP classifiers.

6.3. Language-independent emotion recognition

The emotional speech also varies with the language and cultural differences. Due to the unavailability of labeled data, it is a difficult task to build SER systems for low resource languages. Kamaruddin et al. [207] studied the culture dependency effect for SER. Neumann et al. [178] conducted mono-lingual, multi-lingual, and cross-lingual experiments for emotion recognition. These experiments were evaluated on attentive convolutional neural networks. The authors tried to predict the arousal and valence level for emotions. An improvement in the arousal prediction was observed in multi-lingual experiments due to sufficient data, while there was no improvement in valence prediction. In the cross-lingual experiments, the model was tuned with a limited amount of target data. Again, the results of arousal prediction are comparable with the mono-lingual experiments, whereas there was no improvement in the valence prediction. From these results, it was concluded that the valence prediction is more sensitive to language. Here, spectral and voice quality features may be useful because they are correlated with valence.

Successful language-independent emotion recognition also achieves speaker and text-independent emotion recognition. This is because each corpus belonging to a different language has a different set of speakers and texts. When the model is built with one corpus and tested with other corpora, it is called language-independent (cross-corpus) SER. Domain adaptation techniques are used to adapt a model developed using the source domain data (training data), for which label information is available, to the target domain data (testing data). Such techniques for which no target-labeled data is required are known as unsupervised domain adaptation techniques. The techniques for which both labeled (though in less amount) and unlabeled data are required, are known as semi-supervised domain adaptation techniques.

1. Domain adaptation: The unsupervised domain adaptation approaches in the literature are [27,148,174,208–211]. Yun and Yoo [208] estimated the GMM parameters by margin scaling with a loss function. This gives good generalization capability for both binary and multi-class emotion recognition when the parameters are over-fitted due to mismatch in the training and testing distributions. Deng et al. [148] proposed adaptive-denoising auto-encoder to handle the distribution mismatch between the source and target domains. The goal was achieved using transfer learning, where an auto-encoder was first learned by target domain data in an unsupervised fashion. After that, the auto-encoder was fine-tuned using source domain data to achieve a common feature representation. Zong et al. [209] proposed domain-adaptive least-squares regression (DaLSR) model. While learning, this model used both the source and target domain data. This alleviates the problem of mismatched training and testing distributions by imposing a regularizing restriction to the existing objective function of least squares regression (LSR). Song et al. [210] used the maximum mean discrepancy (MMD) algorithm to minimize the variance between the source and target domain. MMD is used to project the source and target domain data into similar space. This process is called transfer discriminant analysis. MMD uses functions from a kernel Hilbert space as discriminatory functions. The success of MMD depends on the selection of a proper kernel function. The authors showed their improvement over the transfer non-negative factorization (TNF) approach [212]. The accuracy gap was reduced by 22.15% for MMD and 23% for TNF. The accuracy gap was obtained from the system where training and testing were performed on the same database.

Mao et al. [174] proposed emotion-discriminative and domain-invariant feature learning. The proposed system is a multi-tasking framework that simultaneously predicts both emotion and domain. An orthogonal term was defined to minimize the disentangling factors, namely speakers, recording, and noise. A gradient reversal layer was introduced between the domain prediction and feature extraction layers to project the source and target domain in the same distribution. The accuracy was 61.63% for the proposed domain adaptation, while it was 56.52%, 56.90%, and 55.58% for the existing domain adaptation methods shared-hidden-layer autoencoder (SHLA) [213], autoencoder-based unsupervised domain adaptation (AUDA) [148], and sparse autoencoder-based feature transfer learning (SAFTL) [214], respectively. Deng et al. [27] proposed 'universum' auto-encoder for domain adaptation. The universum autoencoder simultaneously predicts the classification and reconstruction loss. The classification loss is the sum of the margin loss for the labeled data and intensive loss for the unlabeled data. The authors showed its robustness against previous domain adaptation methods [213,215,216]. The proposed approach used only 100 labeled training samples from the primarily unlabeled SUSAS database. It achieved unweighted average accuracy (UWA) of 54.8%, a significant improvement using only 100 labeled training samples over the UWA value of 54.1% achieved by SVM when using the entire labeled data set (640 samples).

Abdelwahab and Busso [211] proposed an unsupervised domain adversarial neural network (DANN) that simultaneously minimizes the loss for the domain and class predictions. Between the shared layer and the first layer of the domain classifier, a gradient inversion layer was introduced. During the forward propagation, the gradient inversion layer passes forward the output from the previous layer, while it reverses the sign of the gradient during the back propagation so that the training and testing distributions become similar. The domain adversarial network need not require kernel function, but it needs proper regularization to prevent over-fitting. The accuracy was improved by 6.6% when compared to the model trained with only the source domains.

2. Ensembling: Ensembling [25,26] minimizes the variance between the source and target domains. Abdelwahab and Busso [26] proposed a semi-supervised domain adaptation. It selects features using an ensembling method. The proposed feature selection method chooses features such that the variance between the source and target domain data is minimized. Albornoz and Milone [25] suggested another approach based on ensemble learning in which a separate classifier was proposed for each language. For an unseen language, the decision of emotion was taken by the majority voting and the confidence support by different classifiers. The accuracy was improved by 4.37% for unseen languages when compared with the baseline model.

3. Language-independent features: Here, the goal is to find the features that are invariant to the source and target domain data. Ying and Xue-ying [218] proposed zero crossings with maximal

Table 7
Summary of the language-independent SER techniques.

Reference	Year	Method	Task	Results
Yun and Yoo [208]	2013	GMM parameters are estimated by margin scaling with a novel loss function.	Binary and six class.	Margin scaling using log (MSL) performed better than margin scaling using linear (MSN), which in turn performed better than margin scaling using exponential (MSE).
Deng et al. [148]	2014	Adaptive denoising auto-encoder	2-class classification. Positive vs negative.	It achieved 8.32 % better accuracy than the denoising auto-encoder for cross-corpus SER.
Cairong et al. [217]	2016	Various statistics of spectrogram.	4-class classification.	The new feature subset achieved 8.8 % better accuracy than the traditional feature set.
Zong et al. [209]	2016	Domain-adaptive least-squares regression (DaLSR) model, SVM.	6-class classification.	It performed better than the state-of-the-art transfer learning algorithms for cross-corpus SER.
Mao et al. [174]	2017	DNN, Multi-tasking framework, Gradient reversing layer.	2-class classification. Positive vs negative.	The accuracy was 5.11 %, 4.73% and 6.05% better than the SHLA, AUDA and SAFTL methods, respectively.
Albornoz and Milone [25]	2017	Ensembling approach.	6-class classification.	The accuracy was improved by 4.37% for unseen languages when compared with the baseline model.
Abdelwahab and Busso [26]	2017	Semi-supervised domain adaptation, Feature selection using ensembling method.	4-class classification.	Improvements of 1-3.7% were obtained for different number of samples when compared with the baseline method.
Deng et al. [27]	2017	Universum auto-encoder.	4-class classification.	The universum auto-encoder method achieved 5% better accuracy than the SHLA [213] method.
Song et al. [210]	2017	Maximum mean discrepancy (MMD) algorithm.	6-class classification.	The accuracy gap was reduced by 22.15% for MMD and 23% for TNF.
Ying and Xue-ying [218]	2018	Glottal compensation to zero crossings with maximal teager energy operator (GCZCMT).	3-class classification.	The SER accuracy was improved by 1.08% and 4.72% when compared with MFCC and ZMCT, respectively.
Abdelwahab and Busso [211]	2018	Unsupervised domain adaptation, Domain adversarial neural network (DANN).	Arousal and valence prediction (Regression).	The accuracy was improved by 6.6 % when compared with no adaptation.

Teager energy operator (ZCMT), in which the glottal compensation is performed for language-independent emotion recognition. The SER accuracy was improved by 1.08% and 4.72% when compared with MFCC and ZMCT, respectively. Cairong et al. [217] proposed statistical features of spectrogram. These features are robust against cross-corpus emotion recognition. Further, these features were combined with the acoustic features in a deep belief neural network resulting in an improved SER accuracy. The methods used to compensate for the mismatch between the training and testing distributions are listed in Table 7. Most of the cross-corpus experiments were performed on openSMILE feature sets [219]. These feature sets are hand-crafted. Considering the success of the deep-learning-based features, the above techniques can be combined with deep-learning techniques. These techniques have been applied to both classification and regression tasks. However, there is a need to study each method for both classification and regression task simultaneously.

6.4. Uncontrolled-environment

Various elements in natural data can not be controlled, e.g., noise, codec-effect, and microphone-distance. Most of the SER systems are built with acted databases, where the data is recorded in a controlled environment. The performance of such systems dramatically degrades when tested with natural data. This is because of the mismatched distribution of the training and testing data. Due to the different types of background noises, natural data cannot be directly used for training and testing purposes. Next, we discuss the three common elements associated with uncontrolled data, namely noise, codec-effect, and microphone-distance, and the techniques proposed to deal with them.

A. Noisy-environment: It is important to consider the effect of the background noise for SER to develop suitable solutions for real-world applications, e.g., the case of monitoring threatening calls when the caller is using a public phone on a street, cafeteria, railway station, airport, war place, etc. Most of the SER systems are developed using data that was recorded in a clean environment. In a real-life application, the nature of data can not be controlled. There are different types of background noises present along with the speech. A number of techniques have been proposed for dealing with noises, however, their scope of discussion/analysis in the existing literature has been rather limited to a few types of noises. This motivates the need to explore such techniques for a wider variety of noises. The noisy environments can be divided into the following categories.

1. Additive white Gaussian noise (AWGN): The energy is uniformly distributed across all the frequencies. Such type of noise is generally produced by the fan, cooler, and air-conditioning devices.

2. Cafeteria babble: As the name suggests, such type of noise is typically produced in a cafeteria-like environment. It is characterized by the interference of external voices produced by other speakers. The cafeteria babble may also contain several impulsive noises, e.g., those produced by the clinking of dishes, coughing or laughing. The cafeteria noise exhibits higher power values in the frequency range of 400 Hz to 1 kHz. The frequency components of

noise are widely dispersed throughout the spectrum and vary with time.

3. Street noise: As the name indicates, such noises are generated in the principal or centric avenues of a city. They can be generated by distant vehicles and other remote items, because of which the frequency elements are distributed in the low-frequency region, however, the noise generated by nearby vehicles can produce high-frequency components. The street noise may also contain impulsive sounds produced by the vehicle horns.

6.4.1. Techniques

Now we discuss techniques proposed to deal with the noisy data.

1. Noise removal techniques: Some popular techniques for noise removal are: logMMSE [220], Karnuhen-Loeve transform (KLT) decomposition [221], spectral subtraction [222], and wiener filter [222] for speech enhancement. Vasquez-Correa et al. [221] evaluated logMMSE and KLT decomposition for SER. These techniques need the silence regions of fixed duration for characterizing the background noise. Their experiments did not produce any improvement in FAU-Aibo database due to the absence of such silence regions. The WPT and SSWT-based features improved the results after applying logMMSE algorithm for white, street, and cafeteria noises. Chenchah and Lachiri [223] used speech enhancement techniques such as spectral subtraction, wiener filter, and MMSE for real-world speech signals (airport, babble, and cafeteria). The spectral subtraction and MMSE methods are robust against airport and babble noises for emotion recognition.

2. Robust features: Recently, some works [45,224,225] have focused on extracting robust features that are not affected by noise. Chenchah and Lachiri [224] proposed power normalized cepstral coefficients (PNCC) features and its variants for SER. The PNCC features are robust against the car, train, airport, and babble noises. The authors compared PNCC and its variants with the well-known MFCC and perceptual linear prediction (PLP) coefficients. The PNCC features differ from the MFFC features for the following reasons. First, PNCC simulates cochlea behavior based on equivalent rectangular bandwidth by gammatone auditory filters. Second, PNCC utilizes an extra step that is the subtraction of the medium-magnitude bias from the output of the gammatone auditory filters. The bias level was computed based on the ratio of the arithmetic to geometric mean (AM-GM ratio) of the medium-time magnitude. Finally, the power function non-linearity was achieved with 1/15 exponent.

Mansour and Lachiri [225] showed the robustness of the MFCC-shifted-delta-cepstral (SDC) coefficients features for airport noise. The author claimed that MFCC-SDC outperformed the traditional MFCC and LPCC features in both the clean and noisy environments. The MFCC and MFCC-SDC features work well for anger and neutral emotions in noisy environments. Further, this work was extended by Mansour et al. [45] who showed the robustness of MFCC-SDC for airport, car, train, and babble noises. The MFCC-SDC features are more robust than the conventional features (MFCC, LPCC, and RASTA-PLP) in a noisy environment. Their use overcomes the limitations of the conventional derivation of the short time cepstral features. The MFCC-SDC features are an extension of the delta-cepstral features and are obtained by stacking delta-cepstra computed across multiple frames.

Huang et al. [46] proposed a sub-band based spectral centroid weighted wavelet packet cepstral coefficients (W-WPCC) and showed their robustness against white noise. The wavelet packet transformation (WPT) is an efficient instrument for non-stationary signal analysis. It is used to analyze speech with a WP filter-bank framework based on human auditory perception. The W-WPCC function is calculated by combining the sub-band energy with the sub-band spectral centroids through a weighting system to produce noise-resistant acoustic characteristics. The accuracy was improved by 1.64% for clean improvement when compared with MFCC and it remains consistent for low SNR levels. This work also highlighted the robustness of IW-SVM classifier [46]. The accuracy was consistent for low SNR levels when compared with SVM and GMM.

All the above-proposed noise-robust features used different databases and classifiers. Hence, the accuracy produced by these features cannot be easily compared. Therefore, the accuracy difference between the proposed and existing features is a suitable parameter for comparison. Among all the newly proposed features, the PNCC features improve the accuracy by around 10% over the traditional features. The PNCC features are also robust against low SNR levels.

There are some features [226–228] that are specifically robust to white noise. Chi et al. [226] extracted different rate-scales, from pitch to speaking rate and from formants to harmonics of the joint spectro-temporal modulation of a spectrogram. They showed its robustness against white noise. While using the MFCC+prosody features, the authors obtained a difference of 32.01% in the accuracy values of the clean and noisy environments (0 dB SNR level). The corresponding number for the openSMILE features is 23.08%. However, this difference was significantly reduced to 7.01% for the proposed features, showing the robustness of the proposed features against noise. Albornoz et al. [227] proposed novel bio-inspired features. They experimentally demonstrated the robustness of these features for white noise. The features are obtained by taking the mean of the log-spectrum using the auditory spectrogram. Further, a moving average filter was applied; eight polynomial coefficients and four statistics of the smoothened curve were used as feature vectors. The difference in the accuracy values of the clean and noisy environments (0 dB SNR level) was 47.14% for the MFCC features. However, this difference was significantly reduced to 27.94% for the proposed features. These features alone did not give good accuracy, prompted by which they were combined with prosody features, resulting in an improvement of 2.57% in the accuracy value. Jassim et al. [228] proposed a novel set of neurogram features and combined them with traditional INTERSPEECH-2010 features [229]. They showed its robustness against white noise. Neurogram is a 2-D representation of speech signal. The 64 characteristic frequencies of neurogram are divided into eight blocks. The average of each block is taken as features. In this way, the size of each feature vector is 512 for each utterance. These features are robust against white noise when compared to the individual feature sets. When compared to the clean environment (15 dB SNR), the accuracy was reduced by 12.21% for the noisy environment (-5 dB SNR level) for the MFCC features. However, the corresponding reduction is only 6.80% when using the neurogram features, indicating the superiority of the neurogram features for the noisy environment.

3. Robust model and embedding: The spectrogram representation is quite often used to implement the deep-learning frameworks for SER. Huang et al. [175] studied the different types of filters used in CNN and its effect on noise. They used a CNN-LSTM architecture in which four types of convolution operations (temporal only convolution, spectral only convolution, full-spectrum temporal convolution, and spectral-temporal convolution) were studied on log Mels and MFCCs for SER under the noisy and clean conditions. Under both the conditions, the full-spectrum temporal convolution outperformed the other three types. The SER accuracies using the full-spectrum temporal convolution were 94.58% and 86.21% for the clean and noisy environments, respectively. The corresponding values when using the temporal only convolution are 92.92% (clean) and 84.23% (noisy), spectral only convolution are 91.67% (clean) and 82.73% (noisy), and spectral-temporal convo-

lution are 93.75% (clean) and 84.26% (noisy). The MUSAN [230] corpus was used to mix the music and noise in a clean database.

Satt et al. [168] highlighted the sensitiveness of spectrogram to noise. The authors modified the spectrogram such that the harmonic properties are preserved. The modified spectrogram was shown to be robust for speech emotion recognition in the presence of three music signals and four crowded noises. When compared to the clean environment, the accuracy was reduced by 28.70% for the noisy environment (0 dB SNR level) for the spectrogram representation. However, the corresponding reduction is only 4.70% when using the modified spectrogram, highlighting the robustness of the modified spectrogram in the noisy environment.

There are various techniques to deal with the issues of SER in the noisy environments that are independent of features [24,231–235]. To discover the inherent geometry of an emotional expression, an improved Lipschitz embedding technique was proposed in [231]. This method uses the geodesic distance that embeds the 64-dimensional acoustic features into a six-dimensional space. Both the Gaussian white and sinusoidal noises were added to the speech database at several SNRs. The SER accuracy was improved by 10% in the noisy environments when compared with LDA and PCA. Juszkiewicz [232] proposed the histogram equalization method to reduce the differences between the feature vectors in the clean and noisy environments. During training, the histograms of pitch and MFCC were averaged, which served as a reference for equalization. This method was tested on the ambient noise and sounds of the motors of a robot. When compared to the clean environment, the accuracy was reduced by 56% for the noisy environment (0 dB SNR level) without the histogram equalization. However, the corresponding reduction is only 31% with the histogram equalization, proving the importance of the histogram equalization method in the noisy environment.

Zhao et al. [233] improved sparse representation classifier for robust emotion recognition in the Gaussian white noise. The goal was to discriminatively find a sparse representation of the test samples as a linear combination of the training samples by solving it as an optimization problem. This work improved the accuracy of the classifier by using a weighted sparse representation model based on the maximum likelihood estimate. The SER accuracies using the weighted sparse representation classifier were 71.67% and 63.75% for clean and noisy (0 dB) environments, respectively, while the corresponding values for other classifiers are 66.25% and 56.25% for the SVM, 62.29% and 45.37% for the KNN, and 60.83.75% and 84.26% for the sparse representation classifier. Bashirpour and Geravanchizadeh [234] proposed a method based on the binaural input signal analysis and estimated the emotional auditory mask for recognizing emotions. The binaural signal analyzer segregates the noise from a speech by estimating the emotional mask that removes the most emotionally affected spectro-temporal regions of the segregated target speech. Jing et al. [235] evaluated their proposed noise reduction method on the RECOLA database where the arousal and valence levels were measured with the concordance correlation coefficient (CCC). Noises were reduced by data down-sampling, feature synchronization, and a modified version of graph total variation. This method is robust for both noise and reverberation. Shahin et al. [24] showed the robustness of the GMM-DNN hybrid classifier against the white, siren, and telephone noises. It can be concluded that while both the robust features and feature-independent techniques produce improvements for noisy environments, the robust features may result in reduced performance in a clean environment, however, the performance of feature-independent techniques is largely unaffected in a clean environment. Finally, the methods used to deal with the issues of uncontrolled environments are summarized in Table 8.

B. Codec-data: To make more efficient use of network resources, such as the bandwidth, the codecs compress the speech signals by reducing the bit-rate. Apart from the effects of the background noise, the effect of different telephony codecs and channels also needs to be evaluated. Speech may be recorded by different sources and in different conditions that may vary in terms of factors such as microphones, sampling frequency, and the number of quantization bits. Such recordings may be collected from a call-center, a mobile phone, or from a video-conference using tools like Skype or Google Hangouts. Due to the lowering of bit-rates, some information is removed from the speech signal, leading to a lower accuracy for SER.

Vasquez-Correa et al. [221] studied the effect of codecs in SER. Speech codecs are popular in VoIP and mobile telephone channels. There are seven codecs that are used for speech compression: AMR-NB, AMR-WB, GSM, G.722, G.726, SILK, and Opus [221]. The results were evaluated on the OpenEAR, spectral+noise+NLD, WPT, and SSWT features. It was concluded that the SSWT features are most robust with respect to codec-speech. This is because of the concentrated time-frequency representation of the signal in the narrow-band. SSWT features are also robust to white noise and other perturbations [236]. In another work, Albahri et al. [237] studied the effect of the adaptive multi-rate wideband (AMR-WB) and extended adaptive multi-rate wideband (AMR-WB+) codecs using the MFCCs and glottal features from the time and frequency domain signals. Further, Teager energy operator-perceptual wavelet packet (TEO-PWP) was proposed and its robustness against codec-speech was experimentally demonstrated. The SSWT and TEO-PWP features are more robust for codec-speech than the conventional features. All said, there is still a significant need to develop and explore such features whose accuracy is comparable to the uncompressed features.

C. Microphone-distance: Distant emotion identification (DER) expands the application of voice emotion to the very challenging scenario where the speech signal is collected by a distant microphone. The efficiency of conventional emotion identification systems degrades very quickly and dramatically as the microphone is moved away from the speaker's mouth. This is due to the following reasons: (1) lowering of SNR due to the background noise, (2) voice interference from other speakers, (3) reverberation of sound due to reflections, (4) noise diffusion, and (5) reduced amplitude of the direct sound. The acoustic characteristics of the distant microphone voice signals are therefore not the same as those of the corresponding closed microphone voice signals. The effects of distant microphone are studied in [132,238–240]. Salekin et al. [240] studied 77 LLD features and their delta and delta-delta features (a total 231 features) for DER. The accuracy of SER significantly decreased with the delta and delta-delta features as the microphone distance was increased. For most of the features, other than delta and delta-delta, the relative accuracy reduced about 40% to 50% at a fixed distance of six meters, while the corresponding figures for delta and delta-delta were more than 100%. Finally, 48 LLD features were selected with less than 50% distortion through the various distance measures. These 48 features are the five MFCCs, voice probability, fundamental frequency, zero-crossing rate, eight line spectral pair frequencies, and 32 logarithmic power of Mel-frequency bands. Ahmed et al. [132] also used feature selection assuming that the selected features are less distorted with distance. Initially, the distortion value of each of the 6552 features with respect to its corresponding clean signal feature value was computed. The features were then arranged in a descending order of their distortion. The distorted features were eliminated iteratively (one-by-one) from the train and test sets. A new emotional model with the updated reduced-by-one feature set was constructed and tested on the test set for each iteration, logging the corresponding classification accuracy. The existing works on DER involve pick and chose from the existing SER feature set based on how unaffected they remain as the distance is increased. In general, most of

Table 8
Summary of SER methods in uncontrolled environments.

1. SER in noisy environments				
Refrence	Year	Method	Noise Type	Results
You and Chen [231]	2006	Lipschitz embedding technique.	Gaussian white noise and sinusoidal noise.	The SER accuracy was improved by 10% in the noisy environments when compared with LDA and PCA.
Chi et al. [226]	2012	Statistics of the joint spectro-temporal modulation features.	White and babble.	The accuracy gap (between clean and noisy) was 32.01% for the MFCC+prosody features, while it was only 7.01% for the proposed features.
Vasquez-Correa et al. [221]	2014	logMMSE, KLT.	Babble, street, and AWGN.	logMMSE algorithm improved the results for all the features. WPT and SSWT based features were found to be robust to noise.
Juszkiewicz [232]	2014	Histogram equalization.	Ambient noise, sounds of motors of a robot.	The accuracy gap (between clean and noisy) was 56% without histogram equalization, while it was only 31% with histogram equalization.
Zhao et al. [233]	2014	Sparse representation classifier.	Gaussian white noise.	The improvements in the clean environment were 5.42% and 9.38% for SVM and KNN, respectively, while the corresponding improvements for the noisy environment were 7.5% and 18.38%, respectively.
Mao et al. [82]	2014	CNN is used to find the salient features with a novel objective function.	Gaussian noise.	The accuracy gap (between clean and noisy) was only 8.05%.
Chenchah and Lachiri [223]	2016	Spectral subtraction, wiener filter, and MMSE.	Airport and babble.	Spectral subtraction and MMSE methods are robust against airport and babble noise.
Chenchah and Lachiri [224]	2017	PNCC and its variants features.	Car, train, airport, and babble noises.	The improvement was 10% over the well-known MFCC and perceptual linear prediction (PLP) coefficients.
Mansour and Lachiri [225]	2017	MFCC-shifted-delta-cepstral (SDC) coefficients features.	Airport noise.	MFCC-SDC outperforms the traditional MFCC and LPCC parameters in the clean and noisy environments.
Huang et al. [46]	2017	Sub-band based spectral centroid weighted wavelet packet cepstral coefficients (W-WPCC), the importance of weighted SVM.	White Gaussian noise.	W-WPCC outperforms MFCC by 1.64% in noisy environments.
Jassim et al. [46]	2017	Neurogram features.	White noise.	The accuracy gap (between clean and noisy) was 12.21% for the MFCC features, while it was only 6.80% for the proposed features.
Satt et al. [168]	2017	Modified spectrogram for CNN-LSTM.	Three music signals and four crowd-source noises.	The accuracy gap (between clean and noisy) was 28.70% for the spectrogram, while it was only 4.70% for the modified spectrogram.
Huang et al. [175]	2017	Different types of convolution (CNN-LSTM) operation was studied in noisy environments.	MUSAN corpus is used to mix the music and noise	Full-spectrum temporal convolution showed improvement of about 2% in both clean and noisy environments.
Jing et al. [235]	2018	Graph total variation regularization (GTVR).	Reverberation and additive noises (car, living room, train, and restaurant noises).	The CCC values of arousal and valence prediction are robust for reverberation and additive noise.
Bashirpour and Geravanchizadeh [234]	2018	Binaural signal analyzer.	White Gaussian noise.	It isolates the noise from speech by constructing a speech mask in a noisy environment.
Mansour et al. [45]	2018	MFCC-SDC coefficients.	Airport, car, train, and babble noises.	It is 3.5%, 13.85 % and 6% better than MFCC, LPCC, and RASTA-PLP respectively in the clean environment. The corresponding improvements were 11.15%, 11.78%, and 8.77% in the noisy environment.
Shahin et al. [24]	2019	GMM-DNN hybrid classifier.	White, siren, and telephone noises.	The dominant signal mask provided by the hybrid classifier improved upon the performance by 4.6% and 15.1% over the SVM and MLP, respectively, in the presence of the noisy signals.
2. SER from telephonic (codec) speech				
Reference	Year	Method	Results	
Camilo et al. [221]	2014	SSWT features.	Codecs:-AMR-NB, AMR-WB, GSM, G.722, G.726, SILK, and Opus.	The SSWT features are more robust to codec-speech.

(*continued on next page*)

Table 8 (*continued*)

3. SER from distant speech				
Reference	Year	Method		Results
Salekin et al. [240]	2017	Feature selection based on lesser distortion with distance.	25	The delta, delta-delta features are more sensitive to distance. Five Mel-frequency cepstral coefficients, voice probability, fundamental frequency, zero-crossing rate, eight line spectral pair frequencies, and 32 Logarithmic power of Mel-frequency bands (48) features were found to be less sensitive with distance.
Ahmed et al. [132]	2017	Features are ranked in the order of their distortion.		The accuracy was improved by 15.51% over the algorithm not using the feature selection approach across various rooms and source to microphone distances. Distorted features were eliminated iteratively.

the SER features are based on amplitude and are highly correlated with distance, resulting in performance degradation with increased distance. Hence, there is still a need to extract and explore features that are agnostic to the distance from the source.

6.5. Continuous speech emotion recognition

Until now, all SER systems have been developed and tested for the utterance level SER. Generally, in databases utterances are segmented in advance. However, in a natural environment, people express their speech in a continuous way and their emotions may change with time. In this regard, work has been done [241,242] to detect the points in a speech signal where emotions change, followed by automatic logical decomposition of continuous speech at those points. Even though a few approaches are available to solve this problem of mixed emotion utterances in natural databases, they are not much accurate. This, in turn, is primarily because of the low accuracy in the detection of the points of change in emotions. Hence, there is a need to develop robust techniques to detect the points of change in emotions. Cao et al. [243] proposed a ranking based SVM for mixed emotion utterances (as compared to single emotion utterances). This technique offers two natural advantages. First, it is based on the intuition that each utterance can convey a mixture of possible emotions. Based on this, each utterance is utilized productively to define the dominant emotion by considering the extent to which each emotion is expressed. It also captures emotion-specific information even under speaker-independent training/testing circumstances. The experiments were performed on both acted and spontaneous databases. In both cases, there is a significant improvement in accuracy as compared to the conventional SVM. Happy and disgust emotions are well recognized by the ranking SVM classifiers. This method is used for the testing data where mixed emotion utterances are present.

7. Evaluation metrics

Emotion recognition can be modeled as both classification and regression problems. The evaluation metrics are also different for both types of problems. Accuracy is a popular metric for classification that predicts the true class samples over the total samples. This metric is suitable for a balanced dataset. However, this metric is not suitable for imbalanced dataset because it gives more weight to the class whose samples are more and less weight to the class whose samples are less. This is also called weighted accuracy (WA). Therefore, for imbalance datasets, unweighted accuracy (UWA) is more popular. It is calculated as the average of the individual class accuracies [244]. The UWA is not affected for the imbalanced dataset. There are some other metrics, such as recall, precision and F1-score, which are also used for SER classification problem. Recall measures how many relevant samples are retrieved whereas precision measures how many retrieved samples are relevant. The use of recall and precision depends on application. For example, the diagnostic of critical diseases requires more precision, while the detection of criminal activities require high values for both precision and recall. The high value for both precision and recall can not be achieved simultaneously. Therefore, F1-score [245] – defined as the harmonic mean of the recall and precision values – is used for those applications that require high value for both the recall and precision. The equal error rate (EER) [246] is another measure used for SER that cares for both the true positive rate (TPR) and the false positive rate (FPR). EER is a point in a receiver operating characteristic (ROC) curve where the TPR and FPR are equal.

The continuous emotion recognition is modeled as a regression problem. The root mean square error (RMSE), Pearson's correlation coefficient (PR), and concordance correlation coefficient (CCC) metrics are used for continuous emotion recognition [211]. The CCC is the combination of both RMSE and PR, therefore, it is a more suitable metric. RMSE gives information about how far the predicted value deviates from the ground truth value, while PR value gives information about the association between the predicted and ground truth values. Lower RMSE and higher PR and CCC values for arousal, valence and dominance decide which model is better. The CCC is typically more restrictive and has a lower value than the PR. However, it can be a better approach to estimate the quality of regression output in the case of those few situations where there may be an issue with a strong bias in predictions.

8. Discussion and conclusions

This article presented a comprehensive survey of SER in the natural environment. It discussed various issues of SER in natural environment and related solutions to deal with those issues. The new challenging databases are described with their issues in detail.

In this survey, the features are categorized into two broad categories of acoustic and non-acoustic. Further, the acoustic features are categorized into the prosody, spectral, wavelet-based, voice-quality, non-linear, and deep-learning-based features. The non-acoustic features are categorized into linguistic, discourse, face, gesture, and video. The strength of each type of acoustic feature is studied with respect to different emotions. This study helps to select the right features for SER. The scenarios are also presented where the non-acoustic features may be useful in combination with the acoustic features. The feature selection algorithms are discussed with respect to SER. The related discussion in this paper will help in selecting suitable feature selection algorithms.

There are numerous deep-learning frameworks that superseded the accuracy of the conventional methods; these are discussed. In this survey, the deep-learning-based issues in a natural environment and their solutions in the literature are discussed. However, the issues are not completely resolved, thereby motivating further research. Most of the deep-learning methods used spectrogram as an input. The spectrogram gives better results than the conventional features. It is an open question for future research: is there at all a need for feature engineering in deep-learning? This survey also emphasizes on the use of hierarchical classifiers. It can work better when at each node a different subset of features is selected

for a particular group of emotions. For example, pitch related features are better for classifying happy and sad emotions.

Specifically, this survey discusses emotion recognition in the mismatched (with respect to speaker, text, and language-independent emotion recognition) and uncontrolled (with respect to noisy, codec, and distant speech) environments. Various approaches with their advantages and disadvantages are discussed to deal with the mismatched environments.

The RSS and bi-spectral features have shown more robustness than the conventional features for the speaker-independent emotion recognition. This approach does not require the speaker information to minimize the speaker effect on emotion recognition. The speaker normalization methods also do not require the speaker information to minimize the variance among speakers. The z-normalization and vocal-tract normalization methods have been used for speaker normalization. The z-normalization method is simple and easy. MAP and MLLR are used for speaker adaptation where the model is re-estimated for each speaker. MLLR is an unsupervised technique that does not require labeled emotional data for re-estimating the model, thus, it is more popular. Both the MAP and MLR methods require the speaker information for the speaker adaptation. This creates a problem for the speakers in the testing set, where the number of samples is very less – also known as the cold start problem. There are some other methods such as ensembling, covariance-shift, and discriminative training for speaker-independent emotion recognition. The objective of these methods is to minimize the variance among speakers and to maximize the variance among emotions.

The RSS, adaptive TEO, pitch-based features, and glottal spectrogram for deep-learning are more robust than the conventional features with respect to text-independent emotion recognition. The whitening transform and functional PCA methods are used to transform the features for feature selection. The whitening transform method is more popular due to its simplicity. The ensembling of generative-discriminative (GMM-DNN) classifiers reduces the variance produced due to the different text in the training and testing data.

For language-independent SER, the glottal compensation to zero-crossings with maximal Teager energy operator (GCZCMT) features are more robust than the conventional features. The ensembling methods are also robust for this purpose. Recently, cross-corpus experiments are more popular because they inherently create the experiment speaker-independent, text-independent and language-independent. The domain-adversarial network provided a satisfactory result for the cross-corpus experiments. There is still scope for the researcher to develop a methodology that could work for cross-corpus experiments.

The performance degrades in natural databases due to uncontrolled environments (i.e., various types of noises, the effect of codecs, and microphone distance). Various techniques are discussed to address the effect of noise. Some speech enhancement techniques have also been evaluated for SER. These techniques require more time due to the extra preprocessing steps required to remove noise from the speech signal. Therefore, most of the works tried to develop features that are robust to background noise. The power normalized cepstral coefficients (PNCC), MFCC-shifted-delta-cepstral (SDC) coefficients, and weighted wavelet packet cepstral coefficients (W-WPCC) features are robust to different types of noises in SER. The speech enhancement techniques and robust features reduce the recognition accuracy for a clean environment. There is a scope for researchers to develop features that work for both clean and noisy environments. Due to the popularity of spectrogram in the case of the deep-learning frameworks, it has also been explored for the noisy environments. We also discussed some techniques that are independent of features and work well for the noisy environments. In general, such techniques work well for both the clean and noisy environments. However, there is still a lot of scope for researchers to develop and explore the deep-learning frameworks in noisy environments.

The codec-effect can be reduced by proper extraction of features. The accuracy of distant SER is enhanced by selecting features that are less affected by the microphone distance. There is still a lot of scope for researchers to find the features that would be robust against microphone distance. The current emerging area of continuous SER, where an utterance is not segmented in advance, as happens in natural environments, is also discussed and has future scope for research.

Declaration of competing interest

The authors certify that they have no competing interest.

Acknowledgments

Akshay Deepak has been awarded Young Faculty Research Fellowship (YFRF) of Visvesvaraya PhD Programme of Ministry of Electronics & Information Technology, MeitY, Government of India. In this regard, he would like to acknowledge that this publication is an outcome of the R&D work undertaken in the project under the Visvesvaraya PhD Scheme of Ministry of Electronics & Information Technology, Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia).

References

[1] M.D. Pell, L. Monetta, S. Paulmann, S.A. Kotz, Recognizing emotions in a foreign language, J. Nonverbal Behav. 33 (2) (2009) 107–120.
[2] R. Nakatsu, J. Nicholson, N. Tosa, Emotion recognition and its application to computer agents with spontaneous interactive capabilities, Knowl.-Based Syst. 13 (7) (2000) 497–504.
[3] D. Ververidis, C. Kotropoulos, A state of the art review on emotional speech databases, in: Proceedings of 1st Richmedia Conference, Citeseer, 2003, pp. 109–119.
[4] T. Sagar, Characterisation and synthesis of emotions in speech using prosodic features, Ph.D. thesis, Master's thesis, Dept. of Electronics and Communications Engineering, Indian Institute of Technology Guwahati, 2007.
[5] C.M. Lee, S.S. Narayanan, Toward detecting emotions in spoken dialogs, IEEE Trans. Speech Audio Process. 13 (2) (2005) 293–303.
[6] K.E.B. Ooi, L.-S.A. Low, M. Lech, N. Allen, Early prediction of major depression in adolescents using glottal wave characteristics and teager energy parameters, in: Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, IEEE, 2012, pp. 4613–4616.
[7] L.-S.A. Low, N.C. Maddage, M. Lech, L.B. Sheeber, N.B. Allen, Detection of clinical depression in adolescents' speech during family interactions, IEEE Trans. Biomed. Eng. 58 (3) (2011) 574–586.
[8] Y. Yang, C. Fairbairn, J.F. Cohn, Detecting depression severity from vocal prosody, IEEE Trans. Affect. Comput. 4 (2) (2013) 142–150.
[9] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, T. Ehrette, Fear-type emotion recognition for future audio-based surveillance systems, Speech Commun. 50 (6) (2008) 487–503.
[10] B. Schuller, A. Batliner, S. Steidl, D. Seppi, Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge, Speech Commun. 53 (9–10) (2011) 1062–1087.
[11] M. El Ayadi, M.S. Kamel, F. Karray, Survey on speech emotion recognition: features, classification schemes, and databases, Pattern Recognit. 44 (3) (2011) 572–587.
[12] S.G. Koolagudi, K.S. Rao, Emotion recognition from speech: a review, Int. J. Speech Technol. 15 (2) (2012) 99–117.
[13] C.-N. Anagnostopoulos, T. Iliou, I. Giannoukos, Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011, Artif. Intell. Rev. 43 (2) (2015) 155–177.
[14] M. Swain, A. Routray, P. Kabisatpathy, Databases, features and classifiers for speech emotion recognition: a review, Int. J. Speech Technol. 21 (1) (2018) 93–120.
[15] M.B. Mustafa, M.A. Yusoof, Z.M. Don, M. Malekzadeh, Speech emotion recognition research: an analysis of research focus, Int. J. Speech Technol. 21 (1) (2018) 137–156.
[16] B.W. Schuller, Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends, Commun. ACM 61 (5) (2018) 90–99.

[17] M.B. Akçay, K. Oğuz, Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers, Speech Commun. 116 (2020) 56–76.
[18] E.H. Kim, K.H. Hyun, S.H. Kim, Y.K. Kwak, Improved emotion recognition with a novel speaker-independent feature, IEEE/ASME Trans. Mechatron. 14 (3) (2009) 317–325.
[19] C. Yogesh, M. Hariharan, R. Yuvaraj, R. Ngadiran, S. Yaacob, K. Polat, et al., Bispectral features and mean shift clustering for stress and emotion recognition from natural speech, Comput. Electr. Eng. 62 (2017) 676–691.
[20] Y. Sun, G. Wen, Emotion recognition using semi-supervised feature selection with speaker normalization, Int. J. Speech Technol. 18 (3) (2015) 317–331.
[21] C. Wu, C. Huang, H. Chen, Text-independent speech emotion recognition using frequency adaptive features, Multimed. Tools Appl. 77 (18) (2018) 24353–24363.
[22] S. Ghosh, E. Laksana, L.-P. Morency, S. Scherer, Representation learning for speech emotion recognition, in: Interspeech, 2016, pp. 3603–3607.
[23] S. Mariooryad, C. Busso, Compensating for speaker or lexical variabilities in speech for emotion recognition, Speech Commun. 57 (2014) 1–12.
[24] I. Shahin, A.B. Nassif, S. Hamsa, Emotion recognition using hybrid Gaussian mixture model and deep neural network, IEEE Access (2019).
[25] E.M. Albornoz, D.H. Milone, Emotion recognition in never-seen languages using a novel ensemble method with emotion profiles, IEEE Trans. Affect. Comput. 8 (1) (2015) 43–53.
[26] M. Abdelwahab, C. Busso, Ensemble feature selection for domain adaptation in speech emotion recognition, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017, pp. 5000–5004.
[27] J. Deng, X. Xu, Z. Zhang, S. Frühholz, B. Schuller, Universum autoencoder-based domain adaptation for speech emotion recognition, IEEE Signal Process. Lett. 24 (4) (2017) 500–504.
[28] S.G. Koolagudi, S.R. Krothapalli, Emotion recognition from speech using subsyllabic and pitch synchronous spectral features, Int. J. Speech Technol. 15 (4) (2012) 495–511.
[29] A. Dhall, R. Goecke, S. Lucey, T. Gedeon, et al., Collecting large, richly annotated facial-expression databases from movies, IEEE Multimed. 19 (3) (2012) 34–41.
[30] Y. Li, J. Tao, L. Chao, W. Bao, Y. Liu, Cheavd: a Chinese natural emotional audio–visual database, J. Ambient Intell. Humaniz. Comput. 8 (6) (2017) 913–924.
[31] F. Ringeval, A. Sonderegger, J. Sauer, D. Lalanne, Introducing the recola multimodal corpus of remote collaborative and affective interactions, in: Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on, IEEE, 2013, pp. 1–8.
[32] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, Iemocap: interactive emotional dyadic motion capture database, Lang. Resour. Eval. 42 (4) (2008) 335.
[33] C.-C. Lee, E. Mower, C. Busso, S. Lee, S. Narayanan, Emotion recognition using a hierarchical binary decision tree approach, Speech Commun. 53 (9–10) (2011) 1162–1171.
[34] S. Deb, S. Dandapat, Emotion classification using segmentation of vowel-like and non-vowel-like regions, IEEE Trans. Affect. Comput. (2017).
[35] S. Deb, S. Dandapat, Multiscale amplitude feature and significance of enhanced vocal tract information for emotion classification, IEEE Trans. Cybern. 49 (3) (2018) 802–815.
[36] A. Tawari, M.M. Trivedi, Speech emotion analysis: exploring the role of context, IEEE Trans. Multimed. 12 (6) (2010) 502–509.
[37] D. Ververidis, C. Kotropoulos, A review of emotional speech databases, in: Proc. Panhellenic Conference on Informatics (PCI), vol. 2003, 2003, pp. 560–574.
[38] A. Mencattini, E. Martinelli, G. Costantini, M. Todisco, B. Basile, M. Bozzali, C. Di Natale, Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure, Knowl.-Based Syst. 63 (2014) 68–81.
[39] S. Steidl, Automatic Classification of Emotion Related User States in Spontaneous Children's Speech, University of Erlangen-Nuremberg, Erlangen, Germany, 2009.
[40] J.H. Hansen, S.E. Bou-Ghazale, Getting started with susas: a speech under simulated and actual stress database, in: Fifth European Conference on Speech Communication and Technology, 1997.
[41] R.L. Diehl, Acoustic and auditory phonetics: the adaptive design of speech sound systems, Philos. Trans. R. Soc. B, Biol. Sci. 363 (1493) (2008) 965–978.
[42] K.S. Rao, S.G. Koolagudi, Robust emotion recognition using sentence, word and syllable level prosodic features, in: Robust Emotion Recognition Using Spectral and Prosodic Features, Springer, 2013, pp. 47–69.
[43] S. Mirsamadi, E. Barsoum, C. Zhang, Automatic speech emotion recognition using recurrent neural networks with local attention, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017, pp. 2227–2231.
[44] B. Stasiak, K. Rychlicki-Kicior, Fundamental frequency extraction in speech emotion recognition, in: International Conference on Multimedia Communications, Services and Security, Springer, 2012, pp. 292–303.
[45] A. Mansour, F. Chenchah, Z. Lachiri, Emotional speaker recognition in real life conditions using multiple descriptors and i-vector speaker modeling technique, Multimed. Tools Appl. (2018) 1–18.
[46] Y. Huang, W. Ao, G. Zhang, Novel sub-band spectral centroid weighted wavelet packet features with importance-weighted support vector machines for robust speech emotion recognition, Wirel. Pers. Commun. 95 (3) (2017) 2223–2238.
[47] Z. Zong, H. Li, Q. Wang, Multi-channel auto-encoder for speech emotion recognition, arXiv preprint, arXiv:1810.10662, 2018.
[48] C. Busso, S. Lee, S. Narayanan, Analysis of emotionally salient aspects of fundamental frequency for emotion detection, IEEE Trans. Audio Speech Lang. Process. 17 (4) (2009) 582–596.
[49] S. Wu, T.H. Falk, W.-Y. Chan, Automatic speech emotion recognition using modulation spectral features, Speech Commun. 53 (5) (2011) 768–785.
[50] O.-W. Kwon, K. Chan, J. Hao, T.-W. Lee, Emotion recognition by speech signals, in: Eighth European Conference on Speech Communication and Technology, 2003.
[51] N. Sugan, N.S.S. Srinivas, L.S. Kumar, M.K. Nath, A. Kanhe, Speech emotion recognition using cepstral features extracted with novel triangular filter banks based on bark and erb frequency scales, Digit. Signal Process. (2020) 102763.
[52] N. Sato, Y. Obuchi, Emotion recognition using mel-frequency cepstral coefficients, Inf. Media Technol. 2 (3) (2007) 835–848.
[53] Y. Pan, P. Shen, L. Shen, Speech emotion recognition using support vector machine, Int. J. Smart Home 6 (2) (2012) 101–108.
[54] M. Glodek, S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kächele, M. Schmidt, H. Neumann, G. Palm, et al., Multiple classifier systems for the classification of audio-visual emotional states, in: International Conference on Affective Computing and Intelligent Interaction, Springer, 2011, pp. 359–368.
[55] N. Dave, Feature extraction methods lpc, plp and mfcc in speech recognition, Int. J. Adv. Res. Eng. Technol. 1 (6) (2013) 1–4.
[56] K. Wang, N. An, B.N. Li, Y. Zhang, L. Li, Speech emotion recognition using Fourier parameters, IEEE Trans. Affect. Comput. 6 (1) (2015) 69–75.
[57] H. Tao, R. Liang, C. Zha, X. Zhang, L. Zhao, Spectral features based on local hu moments of Gabor spectrograms for speech emotion recognition, IEICE Trans. Inf. Syst. 99 (8) (2016) 2186–2189.
[58] S. Ramamohan, S. Dandapat, Sinusoidal model-based analysis and classification of stressed speech, IEEE Trans. Audio Speech Lang. Process. 14 (3) (2006) 737–746.
[59] R. Böck, O. Egorow, I. Siegert, A. Wendemuth, Comparative study on normalisation in emotion recognition from speech, in: International Conference on Intelligent Human Computer Interaction, Springer, 2017, pp. 189–201.
[60] I. Luengo, E. Navas, I. Hernáez, Feature analysis and evaluation for automatic emotion identification in speech, IEEE Trans. Multimed. 12 (6) (2010) 490–501.
[61] J. Deng, X. Xu, Z. Zhang, S. Frühholz, B. Schuller, Exploitation of phase-based features for whispered speech emotion recognition, IEEE Access 4 (2016) 4299–4309.
[62] R. Sharma, L. Vignolo, G. Schlotthauer, M.A. Colominas, H.L. Rufiner, S. Prasanna, Empirical mode decomposition for adaptive am-fm analysis of speech: a review, Speech Commun. 88 (2017) 39–64.
[63] E. Ramdinmawii, V.K. Mittal, Discriminating between high-arousal and low-arousal emotional states of mind using acoustic analysis, 2018.
[64] L. He, M. Lech, J. Zhang, X. Ren, L. Deng, Study of wavelet packet energy entropy for emotion classification in speech and glottal signals, in: Fifth International Conference on Digital Image Processing (ICDIP 2013), vol. 8878, International Society for Optics and Photonics, 2013, 887834.
[65] S.-H. Chen, J.-F. Wang, Speech enhancement using perceptual wavelet packet decomposition and teager energy operator, J. VLSI Signal Process. Syst. Signal Image Video Technol. 36 (2–3) (2004) 125–139.
[66] I. Daubechies, A nonlinear squeezing of the continuous wavelet transform based on auditory nerve models, Wavelets Med. Biol. (1996) 527–546.
[67] Y. Huang, A. Wu, G. Zhang, Y. Li, Extraction of adaptive wavelet packet filterbank-based acoustic feature for speech emotion recognition, IET Signal Process. 9 (4) (2015) 341–348.
[68] H.K. Palo, M.N. Mohanty, Wavelet based feature combination for recognition of emotions, Ain Shams Eng. J. (2017).
[69] J. Kim, A. Toutios, S. Lee, S.S. Narayanan, A kinematic study of critical and non-critical articulators in emotional speech production, J. Acoust. Soc. Am. 137 (3) (2015) 1411–1429.
[70] C. Gobl, A.N. Chasaide, The role of voice quality in communicating emotion, mood and attitude, Speech Commun. 40 (1–2) (2003) 189–212.
[71] P. Gangamohan, S.R. Kadiri, S.V. Gangashetty, B. Yegnanarayana, Excitation source features for discrimination of anger and happy emotions, in: INTERSPEECH, 2014, pp. 1253–1257.
[72] J. Přibil, A. Přibilová, Evaluation of influence of spectral and prosodic features on gmm classification of Czech and Slovak emotional speech, EURASIP J. Audio Speech Music Process. 2013 (1) (2013) 8.
[73] S.R. Krothapalli, S.G. Koolagudi, Characterization and recognition of emotions from speech using excitation source information, Int. J. Speech Technol. 16 (2) (2013) 181–201.

[74] L. Zao, D. Cavalcante, R. Coelho, Time-frequency feature and ams-gmm mask for acoustic emotion classification, IEEE Signal Process. Lett. 21 (5) (2014) 620–624.
[75] E. Yumoto, W.J. Gould, T. Baer, Harmonics-to-noise ratio as an index of the degree of hoarseness, J. Acoust. Soc. Am. 71 (6) (1982) 1544–1550.
[76] H. Kasuya, S. Ogawa, K. Mashima, S. Ebihara, Normalized noise energy as an acoustic measure to evaluate pathologic voice, J. Acoust. Soc. Am. 80 (5) (1986) 1329–1334.
[77] D. Michaelis, T. Gramss, H.W. Strube, Glottal-to-noise excitation ratio–a new measure for describing pathological voices, Acta Acust. Acust. 83 (4) (1997) 700–706.
[78] G. Tamulevičius, R. Karbauskaitė, G. Dzemyda, Selection of fractal dimension features for speech emotion classification, in: 2017 Open Conference of Electrical, Electronic and Information Sciences (eStream), IEEE, 2017, pp. 1–4.
[79] X. Mao, L. Chen, Speech emotion recognition based on parametric filter and fractal dimension, IEICE Trans. Inf. Syst. 93 (8) (2010) 2324–2326.
[80] D.A. Cairns, J.H. Hansen, Nonlinear analysis and classification of speech under stressed conditions, J. Acoust. Soc. Am. 96 (6) (1994) 3392–3400.
[81] A.M. Badshah, N. Rahim, N. Ullah, J. Ahmad, K. Muhammad, M.Y. Lee, S. Kwon, S.W. Baik, Deep features-based speech emotion recognition for smart affective services, Multimed. Tools Appl. 78 (5) (2019) 5571–5589.
[82] Q. Mao, M. Dong, Z. Huang, Y. Zhan, Learning salient features for speech emotion recognition using convolutional neural networks, IEEE Trans. Multimed. 16 (8) (2014) 2203–2213.
[83] R. Liang, H. Tao, G. Tang, Q. Wang, L. Zhao, A salient feature extraction algorithm for speech emotion recognition, IEICE Trans. Inf. Syst. 98 (9) (2015) 1715–1718.
[84] T. Özseven, Investigation of the effect of spectrogram images and different texture analysis methods on speech emotion recognition, Appl. Acoust. 142 (2018) 70–77.
[85] M. Chen, X. He, J. Yang, H. Zhang, 3-d convolutional recurrent neural networks with attention model for speech emotion recognition, IEEE Signal Process. Lett. 25 (10) (2018) 1440–1444.
[86] L. Sun, J. Chen, K. Xie, T. Gu, Deep and shallow features fusion based on deep convolutional neural network for speech emotion recognition, Int. J. Speech Technol. 21 (4) (2018) 931–940.
[87] K. Laskowski, Contrasting emotion-bearing laughter types in multiparticipant vocal activity detection for meetings, in: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2009, pp. 4765–4768.
[88] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, J. Vepa, Speech emotion recognition using spectrogram & phoneme embedding, Proc. Interspeech 2018 (2018) 3688–3692.
[89] A.I. Goldman, C.S. Sripada, Simulationist models of face-based emotion recognition, Cognition 94 (3) (2005) 193–213.
[90] A. Haag, S. Goronzy, P. Schaich, J. Williams, Emotion recognition using bio-sensors: first steps towards an automatic system, in: ADS, Springer, 2004, pp. 36–48.
[91] M. Egger, M. Ley, S. Hanke, Emotion recognition from physiological signal analysis: a review, Electron. Notes Theor. Comput. Sci. 343 (2019) 35–55.
[92] Y. Jiang, W. Li, M.S. Hossain, M. Chen, A. Alelaiwi, M. Al-Hammadi, A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition, Inf. Fusion (2019).
[93] L.S. Chen, T.S. Huang, T. Miyasato, R. Nakatsu, Multimodal human emotion/expression recognition, in: Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, IEEE, 1998, pp. 366–371.
[94] V. Rozgić, S. Ananthakrishnan, S. Saleem, R. Kumar, R. Prasad, Ensemble of svm trees for multimodal emotion recognition, in: Proceedings of the 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, IEEE, 2012, pp. 1–4.
[95] P. Tzirakis, G. Trigeorgis, M.A. Nicolaou, B.W. Schuller, S. Zafeiriou, End-to-end multimodal emotion recognition using deep neural networks, IEEE J. Sel. Top. Signal Process. 11 (8) (2017) 1301–1309.
[96] S. Jing, X. Mao, L. Chen, Prominence features: effective emotional features for speech emotion recognition, Digit. Signal Process. 72 (2018) 216–231.
[97] G. Castellano, L. Kessous, G. Caridakis, Emotion recognition through multiple modalities: face, body gesture, speech, in: Affect and Emotion in Human-Computer Interaction, Springer, 2008, pp. 92–103.
[98] T. Polzehl, A. Schmitt, F. Metze, M. Wagner, Anger recognition in speech using acoustic and linguistic cues, Speech Commun. 53 (9–10) (2011) 1198–1209.
[99] N. Majumder, S. Poria, A. Gelbukh, E. Cambria, Deep learning-based document modeling for personality detection from text, IEEE Intell. Syst. 32 (2) (2017) 74–79.
[100] D. Griol, J.M. Molina, Z. Callejas, Combining speech-based and linguistic classifiers to recognize emotion in user spoken utterances, Neurocomputing (2017).
[101] S. Klaylat, Z. Osman, L. Hamandi, R. Zantout, Emotion recognition in Arabic speech, Analog Integr. Circuits Signal Process. 96 (2) (2018) 337–351.
[102] S. Planet, I. Iriondo, Children's emotion recognition from spontaneous speech using a reduced set of acoustic and linguistic features, Cogn. Comput. 5 (4) (2013) 526–532.
[103] B. Schuller, Recognizing affect from linguistic information in 3d continuous space, IEEE Trans. Affect. Comput. 2 (4) (2011) 192–205.
[104] M. Muszynski, L. Tian, C. Lai, J. Moore, T. Kostoulas, P. Lombardo, T. Pun, G. Chanel, Recognizing induced emotions of movie audiences from multimodal information, IEEE Trans. Affect. Comput. (2019).
[105] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, E. Cambria, Dialoguernn: an attentive rnn for emotion detection in conversations, arXiv preprint, arXiv:1811.00405, 2018.
[106] L.C. De Silva, P.C. Ng, Bimodal emotion recognition, in: Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), IEEE, 2000, pp. 332–335.
[107] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, S. Narayanan, Analysis of emotion recognition using facial expressions, speech and multimodal information, in: Proceedings of the 6th International Conference on Multimodal Interfaces, ACM, 2004, pp. 205–211.
[108] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, S. Narayanan, Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling, in: Proc. INTERSPEECH 2010, Makuhari, Japan, 2010, pp. 2362–2365.
[109] L.C. De Silva, T. Miyasato, R. Nakatsu, Facial emotion recognition using multi-modal information, in: Proceedings of ICICS, 1997 International Conference on Information, Communications and Signal Processing. Theme: Trends in Information Systems Engineering and Wireless Multimedia Communications, Cat, vol. 1, IEEE, 1997, pp. 397–401.
[110] G. Aguilar, V. Rozgić, W. Wang, C. Wang, Multimodal and multi-view models for emotion recognition, arXiv preprint, arXiv:1906.10198, 2019.
[111] S. Tischer, Method and system for customizing voice translation of text to speech, uS Patent 7,483,832, Jan. 27 2009.
[112] L. Guan, Y. He, S.-Y. Kung, Multimedia Image and Video Processing, CRC Press, 2012.
[113] D. Zhang, S. Chen, Z.-H. Zhou, Constraint score: a new filter method for feature selection with pairwise constraints, Pattern Recognit. 41 (5) (2008) 1440–1451.
[114] S. Maldonado, R. Weber, A wrapper method for feature selection using support vector machines, Inf. Sci. 179 (13) (2009) 2208–2217.
[115] T.N. Lal, O. Chapelle, J. Weston, A. Elisseeff, Embedded methods, in: Feature Extraction, Springer, 2006, pp. 137–165.
[116] T. Pfister, P. Robinson, Real-time recognition of affective states from nonverbal features of speech and its application for public speaking skill analysis, IEEE Trans. Affect. Comput. 2 (2) (2011) 66–78.
[117] D. Gharavian, M. Sheikhan, F. Ashoftedel, Emotion recognition improvement using normalized formant supplementary features by hybrid of dtw-mlp-gmm model, Neural Comput. Appl. 22 (6) (2013) 1181–1191.
[118] L. Chen, X. Mao, Y. Xue, L.L. Cheng, Speech emotion recognition: features and classification models, Digit. Signal Process. 22 (6) (2012) 1154–1160.
[119] L. Sun, S. Fu, F. Wang, Decision tree svm model with Fisher feature selection for speech emotion recognition, EURASIP J. Audio Speech Music Process. 2019 (1) (2019) 2.
[120] T. Özseven, A novel feature selection method for speech emotion recognition, Appl. Acoust. 146 (2019) 320–326.
[121] Z.-T. Liu, M. Wu, W.-H. Cao, J.-W. Mao, J.-P. Xu, G.-Z. Tan, Speech emotion recognition based on feature selection and extreme learning machine decision tree, Neurocomputing 273 (2018) 271–280.
[122] D. Gharavian, M. Sheikhan, A. Nazerieh, S. Garoucy, Speech emotion recognition using fcbf feature selection method and ga-optimized fuzzy artmap neural network, Neural Comput. Appl. 21 (8) (2012) 2115–2126.
[123] H. Pérez-Espinosa, C.A. Reyes-García, L. Villaseñor-Pineda, Acoustic feature selection and classification of emotions in speech using a 3d continuous emotion model, Biomed. Signal Process. Control 7 (1) (2012) 79–87.
[124] X. Jiang, K. Xia, L. Wang, Y. Lin, Reordering features with weights fusion in multiclass and multiple-kernel speech emotion recognition, J. Electr. Comput. Eng. (2017) 2017.
[125] S. Demircan, H. Kahramanli, Application of fuzzy c-means clustering algorithm to spectral features for emotion classification from speech, Neural Comput. Appl. 29 (8) (2018) 59–66.
[126] M. You, C. Chen, J. Bu, J. Liu, J. Tao, A hierarchical framework for speech emotion recognition, in: 2006 IEEE International Symposium on Industrial Electronics, vol. 1, IEEE, 2006, pp. 515–519.
[127] C.S. Ooi, K.P. Seng, L.-M. Ang, L.W. Chew, A new approach of audio emotion recognition, Expert Syst. Appl. 41 (13) (2014) 5858–5869.
[128] E. Väyrynen, J. Kortelainen, T. Seppänen, Classifier-based learning of nonlinear feature manifold for visualization of emotional speech prosody, IEEE Trans. Affect. Comput. 4 (1) (2012) 47–56.
[129] W. Zheng, M. Xin, X. Wang, B. Wang, A novel speech emotion recognition method via incomplete sparse least square regression, IEEE Signal Process. Lett. 21 (5) (2014) 569–572.
[130] J. Yan, X. Wang, W. Gu, L. Ma, Speech emotion recognition based on sparse representation, Arch. Acoust. 38 (4) (2013) 465–470.
[131] S. Sahu, R. Gupta, G. Sivaraman, W. AbdAlmageed, C. Espy-Wilson, Adversarial auto-encoders for speech based emotion recognition, arXiv preprint, arXiv:1806.02146, 2018.

[132] M.Y. Ahmed, Z. Chen, E. Fass, J. Stankovic, Real time distant speech emotion recognition in indoor environments, in: Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, ACM, 2017, pp. 215–224.
[133] S. Haq, P.J. Jackson, Multimodal emotion recognition, in: Machine Audition: Principles, Algorithms and Systems, IGI Global, 2011, pp. 398–423.
[134] Y. Attabi, P. Dumouchel, Anchor models for emotion recognition from speech, IEEE Trans. Affect. Comput. 4 (3) (2013) 280–290.
[135] S. Ntalampiras, N. Fakotakis, Modeling the temporal evolution of acoustic parameters for speech emotion recognition, IEEE Trans. Affect. Comput. 3 (1) (2012) 116–125.
[136] V. Sethu, E. Ambikairajah, J. Epps, On the use of speech parameter contours for emotion recognition, EURASIP J. Audio Speech Music Process. 2013 (1) (2013) 19.
[137] L. Li, Y. Zhao, D. Jiang, Y. Zhang, F. Wang, I. Gonzalez, E. Valentin, H. Sahli, Hybrid deep neural network–hidden Markov model (dnn-hmm) based speech emotion recognition, in: Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on, IEEE, 2013, pp. 312–317.
[138] M. Fahad, J. Yadav, G. Pradhan, A. Deepak, et al., Dnn-hmm based speaker adaptive emotion recognition using proposed epoch and mfcc features, arXiv preprint, arXiv:1806.00984, 2018.
[139] B. Schuller, G. Rigoll, M. Lang, Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture, in: 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, IEEE, 2004, pp. I–577.
[140] Y.-h. Kao, L.-s. Lee, Feature analysis for emotion recognition from mandarin speech considering the special characteristics of Chinese language, in: Ninth International Conference on Spoken Language Processing, 2006.
[141] Y. Wang, S. Du, Y. Zhan, Adaptive and optimal classification of speech emotion recognition, in: 2008 Fourth International Conference on Natural Computation, vol. 5, IEEE, 2008, pp. 407–411.
[142] S. Zhang, Emotion recognition in Chinese natural speech by combining prosody and voice quality features, in: International Symposium on Neural Networks, Springer, 2008, pp. 457–464.
[143] Y. Zhou, Y. Sun, J. Zhang, Y. Yan, Speech emotion recognition using both spectral and prosodic features, in: 2009 International Conference on Information Engineering and Computer Science, IEEE, 2009, pp. 1–4.
[144] F. Eyben, M. Wöllmer, B. Schuller, Openear—introducing the Munich open-source emotion and affect recognition toolkit, in: Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on, IEEE, 2009, pp. 1–6.
[145] F. Eyben, A. Batliner, B. Schuller, Towards a standard set of acoustic features for the processing of emotion in speech, in: Proceedings of Meetings on Acoustics 159ASA, vol. 9, ASA, 2010, 060006.
[146] M.C. Sezgin, B. Gunsel, G.K. Kurt, Perceptual audio features for emotion detection, EURASIP J. Audio Speech Music Process. 2012 (1) (2012) 16.
[147] E. Yüncü, H. Hacihabiboglu, C. Bozsahin, Automatic speech emotion recognition using auditory models with binary decision tree and svm, in: Pattern Recognition (ICPR), 2014 22nd International Conference on, IEEE, 2014, pp. 773–778.
[148] J. Deng, Z. Zhang, F. Eyben, B. Schuller, Autoencoder-based unsupervised domain adaptation for speech emotion recognition, IEEE Signal Process. Lett. 21 (9) (2014) 1068–1072.
[149] K. Wang, N. An, L. Li, Speech emotion recognition based on wavelet packet coefficient model, in: Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on, IEEE, 2014, pp. 478–482.
[150] R. Xia, J. Deng, B. Schuller, Y. Liu, Modeling gender information for emotion recognition using denoising autoencoder, in: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, IEEE, 2014, pp. 990–994.
[151] Y. Huang, A. Wu, G. Zhang, Y. Li, Speech emotion recognition based on coiflet wavelet packet cepstral coefficients, in: Chinese Conference on Pattern Recognition, Springer, 2014, pp. 436–443.
[152] N. Yang, J. Yuan, Y. Zhou, I. Demirkol, Z. Duan, W. Heinzelman, M. Sturge-Apple, Enhanced multiclass svm with thresholding fusion for speech-based emotion classification, Int. J. Speech Technol. 20 (1) (2017) 27–41.
[153] F. Dellaert, T. Polzin, A. Waibel, Recognizing emotion in speech, in: Spoken Language, 1996, ICSLP 96, Proceedings, Fourth International Conference on, vol. 3, IEEE, 1996, pp. 1970–1973.
[154] Y. Wang, L. Guan, An investigation of speech-based human emotion recognition, in: Multimedia Signal Processing, 2004 IEEE 6th Workshop on, IEEE, 2004, pp. 15–18.
[155] T.-L. Pao, Y.-T. Chen, J.-H. Yeh, W.-Y. Liao, Combining acoustic features for improved emotion recognition in mandarin speech, in: International Conference on Affective Computing and Intelligent Interaction, Springer, 2005, pp. 279–285.
[156] M. Lugger, B. Yang, The relevance of voice quality features in speaker independent emotion recognition, in: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, vol. 4, IEEE, 2007, pp. IV–17.
[157] B. Schuller, S. Steidl, A. Batliner, The interspeech 2009 emotion challenge, in: Tenth Annual Conference of the International Speech Communication Association, 2009.
[158] C.-H. Wu, W.-B. Liang, Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels, IEEE Trans. Affect. Comput. 2 (1) (2010) 10–21.
[159] S. Gharsalli, B. Emile, H. Laurent, X. Desquesnes, Feature selection for emotion recognition based on random forest, in: VISIGRAPP (4: VISAPP), 2016, pp. 610–617.
[160] F. Noroozi, T. Sapiński, D. Kamińska, G. Anbarjafari, Vocal-based emotion recognition using random forests and decision tree, Int. J. Speech Technol. 20 (2) (2017) 239–246.
[161] L. Zheng, Q. Li, H. Ban, S. Liu, Speech emotion recognition based on convolution neural network combined with random forest, in: 2018 Chinese Control and Decision Conference (CCDC), IEEE, 2018, pp. 4143–4147.
[162] S.-H. Wang, H.-T. Li, E.-J. Chang, A.-Y.A. Wu, Entropy-assisted emotion recognition of valence and arousal using xgboost classifier, in: IFIP International Conference on Artificial Intelligence Applications and Innovations, Springer, 2018, pp. 249–260.
[163] A. Iqbal, K. Barua, A real-time emotion recognition from speech using gradient boosting, in: 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), IEEE, 2019, pp. 1–5.
[164] K. Han, D. Yu, I. Tashev, Speech emotion recognition using deep neural network and extreme learning machine, in: Fifteenth Annual Conference of the International Speech Communication Association, 2014.
[165] G.-B. Huang, Q.-Y. Zhu, C.-K. Siew, Extreme learning machine: theory and applications, Neurocomputing 70 (1–3) (2006) 489–501.
[166] J. Lee, I. Tashev, High-level feature representation using recurrent neural network for speech emotion recognition, in: Sixteenth Annual Conference of the International Speech Communication Association, 2015.
[167] A.M. Badshah, J. Ahmad, N. Rahim, S.W. Baik, Speech emotion recognition from spectrograms with deep convolutional neural network, in: 2017 International Conference on Platform Technology and Service (PlatCon), IEEE, 2017, pp. 1–5.
[168] A. Satt, S. Rozenberg, R. Hoory, Efficient emotion recognition from speech using deep learning on spectrograms, Proc. Interspeech 2017 (2017) 1089–1093.
[169] H.M. Fayek, M. Lech, L. Cavedon, Evaluating deep learning architectures for speech emotion recognition, Neural Netw. 92 (2017) 60–68.
[170] W. Zheng, J. Yu, Y. Zou, An experimental study of speech emotion recognition based on deep convolutional neural networks, in: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2015, pp. 827–831.
[171] J. Zhao, X. Mao, L. Chen, Speech emotion recognition using deep 1d & 2d cnn lstm networks, Biomed. Signal Process. Control 47 (2019) 312–323.
[172] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, B. Schuller, Deep neural networks for acoustic emotion recognition: raising the benchmarks, in: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, IEEE, 2011, pp. 5688–5691.
[173] R. Xia, Y. Liu, A multi-task learning framework for emotion recognition using 2d continuous space, IEEE Trans. Affect. Comput. 8 (1) (2015) 3–14.
[174] Q. Mao, G. Xu, W. Xue, J. Gou, Y. Zhan, Learning emotion-discriminative and domain-invariant features for domain adaptation in speech emotion recognition, Speech Commun. 93 (2017) 1–10.
[175] C.-W. Huang, S. Narayanan, et al., Characterizing types of convolution in deep convolutional recurrent neural networks for robust speech emotion recognition, arXiv preprint, arXiv:1706.02901, 2017.
[176] M. Neumann, N.T. Vu, Attentive convolutional neural network based speech emotion recognition: a study on the impact of input features, signal length, and acted speech, arXiv preprint, arXiv:1706.00612, 2017.
[177] P. Li, Y. Song, I. McLoughlin, W. Guo, L. Dai, An attention pooling based representation learning method for speech emotion recognition, Proc. Interspeech 2018 (2018) 3087–3091.
[178] M. Neumann, et al., Cross-lingual and multilingual speech emotion recognition on English and French, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 5769–5773.
[179] K.-Y. Huang, C.-H. Wu, T.-H. Yang, M.-H. Su, J.-H. Chou, Speech emotion recognition using autoencoder bottleneck features and lstm, in: 2016 International Conference on Orange Technologies (ICOT), IEEE, 2016, pp. 1–4.
[180] Z.-w. Huang, W.-t. Xue, Q.-r. Mao, Speech emotion recognition with unsupervised feature learning, Front. Inf. Technol. Electron. Eng. 16 (5) (2015) 358–366.
[181] R. Lotfian, C. Busso, Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings, IEEE Trans. Affect. Comput. (2017).
[182] N. Weißkirchen, R. Bock, A. Wendemuth, Recognition of emotional speech with convolutional neural networks by means of spectral estimates, in: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), IEEE, 2017, pp. 50–55.
[183] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers, B. Schmauch, Cnn+ lstm architecture for speech emotion recognition with data augmentation, arXiv preprint, arXiv:1802.05630, 2018.

[184] I. Siegert, R. Böck, A. Wendemuth, Using a pca-based dataset similarity measure to improve cross-corpus emotion recognition, Comput. Speech Lang. 51 (2018) 1–23.
[185] B. Schuller, D. Arsic, G. Rigoll, M. Wimmer, B. Radig, Audiovisual behavior modeling by combined feature spaces, in: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, vol. 2, IEEE, 2007, pp. II–733.
[186] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, B. Weiss, A database of German emotional speech, in: Ninth European Conference on Speech Communication and Technology, 2005.
[187] O. Martin, I. Kotsia, B. Macq, I. Pitas, The enterface'05 audio-visual emotion database, in: Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on, IEEE, 2006, p. 8.
[188] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, R. Cowie, Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies, in: Proc. 9th Interspeech 2008 Incorp. 12th Australasian Int. Conf. on Speech Science and Technology SST 2008, Brisbane, Australia, 2008, pp. 597–600.
[189] S. Steininger, F. Schiel, O. Dioubina, S. Raubold, Development of user-state conventions for the multimodal corpus in smartkom, in: Proc. Workshop on Multimodal Resources and Multimodal Systems Evaluation, 2002, pp. 33–37.
[190] P. Song, Y. Jin, L. Zhao, M. Xin, Speech emotion recognition using transfer learning, IEICE Trans. Inf. Syst. 97 (9) (2014) 2530–2532.
[191] S. Latif, R. Rana, S. Younis, J. Qadir, J. Epps, Transfer learning for improving speech emotion classification accuracy, arXiv preprint, arXiv:1801.06353, 2018.
[192] D. Tang, J. Zeng, M. Li, An end-to-end deep learning framework for speech emotion recognition of atypical individuals, in: Interspeech, 2018, pp. 162–166.
[193] P.-Y. Shih, C.-P. Chen, H.-M. Wang, Speech emotion recognition with skew-robust neural networks, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017, pp. 2751–2755.
[194] J. Bang, T. Hur, D. Kim, J. Lee, Y. Han, O. Banos, J.-I. Kim, S. Lee, et al., Adaptive data boosting technique for robust personalized speech emotion in emotionally-imbalanced small-sample environments, Sensors 18 (11) (2018) 3744.
[195] J. Chang, X. Zhang, Q. Zhang, Y. Sun, Investigating duration effects of emotional speech stimuli in a tonal language by using event-related potentials, IEEE Access 6 (2018) 13541–13554.
[196] S. Zhang, S. Zhang, T. Huang, W. Gao, Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching, IEEE Trans. Multimed. 20 (6) (2017) 1576–1590.
[197] I.C. Yadav, G. Pradhan, Pitch and noise normalized acoustic feature for children's asr, Digit. Signal Process. (2020) 102922.
[198] J.-H. Yang, J.-w. Hung, A preliminary study of emotion recognition employing adaptive gaussian mixture models with the maximum a posteriori principle, in: 2014 International Conference on Information Science, Electronics and Electrical Engineering, Vol. 3, IEEE, 2014, pp. 1576–1579.
[199] B. Schuller, S. Reiter, R. Muller, M. Al-Hames, M. Lang, G. Rigoll, Speaker independent speech emotion recognition by ensemble classification, in: 2005 IEEE International Conference on Multimedia and Expo, IEEE, 2005, pp. 864–867.
[200] A. Hassan, R. Damper, M. Niranjan, On acoustic emotion recognition: compensating for covariate shift, IEEE Trans. Audio Speech Lang. Process. 21 (7) (2013) 1458–1468.
[201] M. Kockmann, L. Burget, et al., Application of speaker-and language identification state-of-the-art techniques for emotion recognition, Speech Commun. 53 (9–10) (2011) 1172–1185.
[202] C. Busso, A. Metallinou, S.S. Narayanan, Iterative feature normalization for emotional speech detection, in: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, IEEE, 2011, pp. 5692–5695.
[203] C. Busso, S. Mariooryad, A. Metallinou, S. Narayanan, Iterative feature normalization scheme for automatic emotion detection from speech, IEEE Trans. Affect. Comput. 4 (4) (2013) 386–397.
[204] J.-B. Kim, J.-S. Park, Multistage data selection-based unsupervised speaker adaptation for personalized speech emotion recognition, Eng. Appl. Artif. Intell. 52 (2016) 126–134.
[205] S.G. Koolagudi, S. Maity, V.A. Kumar, S. Chakrabarti, K.S. Rao, Iitkgp-sesc: speech database for emotion analysis, in: International Conference on Contemporary Computing, Springer, 2009, pp. 485–492.
[206] J.P. Arias, C. Busso, N.B. Yoma, Shape-based modeling of the fundamental frequency contour for emotion detection in speech, Comput. Speech Lang. 28 (1) (2014) 278–294.
[207] N. Kamaruddin, A. Wahab, C. Quek, Cultural dependency analysis for understanding speech emotion, Expert Syst. Appl. 39 (5) (2012) 5115–5133.
[208] S. Yun, C.D. Yoo, Loss-scaled large-margin Gaussian mixture models for speech emotion classification, IEEE Trans. Audio Speech Lang. Process. 20 (2) (2011) 585–598.
[209] Y. Zong, W. Zheng, T. Zhang, X. Huang, Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression, IEEE Signal Process. Lett. 23 (5) (2016) 585–589.
[210] P. Song, S. Ou, Z. Du, Y. Guo, W. Ma, J. Liu, W. Zheng, Learning corpus-invariant discriminant feature representations for speech emotion recognition, IEICE Trans. Inf. Syst. 100 (5) (2017) 1136–1139.
[211] M. Abdelwahab, C. Busso, Domain adversarial for acoustic emotion recognition, IEEE/ACM Trans. Audio Speech Lang. Process. 26 (12) (2018) 2423–2435.
[212] P. Song, S. Ou, W. Zheng, Y. Jin, L. Zhao, Speech emotion recognition using transfer non-negative matrix factorization, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2016, pp. 5180–5184.
[213] J. Deng, R. Xia, Z. Zhang, Y. Liu, B. Schuller, Introducing shared-hidden-layer autoencoders for transfer learning and their application in acoustic emotion recognition, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2014, pp. 4818–4822.
[214] J. Deng, Z. Zhang, E. Marchi, B. Schuller, Sparse autoencoder-based feature transfer learning for speech emotion recognition, in: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, IEEE, 2013, pp. 511–516.
[215] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, B. Schölkopf, Covariate shift by kernel mean matching, Dataset Shift Mach. Learn. 3 (4) (2009) 5.
[216] T. Kanamori, S. Hido, M. Sugiyama, Efficient direct density ratio estimation for non-stationarity adaptation and outlier detection, in: Advances in Neural Information Processing Systems, 2009, pp. 809–816.
[217] Z. Cairong, Z. Xinran, Z. Cheng, Z. Li, A novel dbn feature fusion model for cross-corpus speech emotion recognition, J. Electr. Comput. Eng. 2016 (2016).
[218] S. Ying, Z. Xue-Ying, Characteristics of human auditory model based on compensation of glottal features in speech emotion recognition, Future Gener. Comput. Syst. 81 (2018) 291–296.
[219] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: the Munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th ACM International Conference on Multimedia, ACM, 2010, pp. 1459–1462.
[220] Y. Ephraim, D. Malah, Speech enhancement using a minimum mean-square error log-spectral amplitude estimator, IEEE Trans. Acoust. Speech Signal Process. 33 (2) (1985) 443–445.
[221] J.C. Vásquez-Correa, N. García, J.F. Vargas-Bonilla, J.R. Orozco-Arroyave, J.D. Arias-Londoño, M.L. Quintero, Evaluation of wavelet measures on automatic detection of emotion in noisy and telephony speech signals, in: 2014 International Carnahan Conference on Security Technology (ICCST), IEEE, 2014, pp. 1–6.
[222] P.C. Loizou, Speech Enhancement: Theory and Practice, CRC Press, 2013.
[223] F. Chenchah, Z. Lachiri, Speech emotion recognition in noisy environment, in: 2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), IEEE, 2016, pp. 788–792.
[224] F. Chenchah, Z. Lachiri, A bio-inspired emotion recognition system under real-life conditions, Appl. Acoust. 115 (2017) 6–14.
[225] A. Mansour, Z. Lachiri, A comparative study in emotional speaker recognition in noisy environment, in: 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), IEEE, 2017, pp. 980–986.
[226] T.-S. Chi, L.-Y. Yeh, C.-C. Hsu, Robust emotion recognition by spectro-temporal modulation statistic features, J. Ambient Intell. Humaniz. Comput. 3 (1) (2012) 47–60.
[227] E.M. Albornoz, D.H. Milone, H.L. Rufiner, Feature extraction based on bio-inspired model for robust emotion recognition, Soft Comput. 21 (17) (2017) 5145–5158.
[228] W.A. Jassim, R. Paramesran, N. Harte, Speech emotion classification using combined neurogram and interspeech 2010 paralinguistic challenge features, IET Signal Process. 11 (5) (2017) 587–595.
[229] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, S.S. Narayanan, The interspeech 2010 paralinguistic challenge, in: Eleventh Annual Conference of the International Speech Communication Association, 2010.
[230] D. Snyder, G. Chen, D. Povey, Musan: a music, speech, and noise corpus, arXiv preprint, arXiv:1510.08484, 2015.
[231] M. You, C. Chen, J. Bu, J. Liu, J. Tao, Emotion recognition from noisy speech, in: 2006 IEEE International Conference on Multimedia and Expo, IEEE, 2006, pp. 1653–1656.
[232] Ł. Juszkiewicz, Improving noise robustness of speech emotion recognition system, in: Intelligent Distributed Computing VII, Springer, 2014, pp. 223–232.
[233] X. Zhao, S. Zhang, B. Lei, Robust emotion recognition in noisy speech via sparse representation, Neural Comput. Appl. 24 (7–8) (2014) 1539–1553.
[234] M. Bashirpour, M. Geravanchizadeh, Robust emotional speech recognition based on binaural model and emotional auditory mask in noisy environments, EURASIP J. Audio Speech Music Process. 2018 (1) (2018) 9.
[235] S. Jing, X. Mao, L. Chen, M.C. Comes, A. Mencattini, G. Raguso, F. Ringeval, B. Schuller, C. Di Natale, E. Martinelli, A closed-form solution to the graph total variation problem for continuous emotion profiling in noisy environment, Speech Commun. 104 (2018) 66–72.
[236] G. Thakur, E. Brevdo, N.S. Fučkar, H.-T. Wu, The synchrosqueezing algorithm for time-varying spectral analysis: robustness properties and new paleoclimate applications, Signal Process. 93 (5) (2013) 1079–1094.
[237] A. Albahri, M. Lech, E. Cheng, Effect of speech compression on the automatic recognition of emotions, Int. J. Signal Proc. Systems 4 (1) (2016) 55–61.

[238] C. Evers, Blind dereverberation of speech from moving and stationary speakers using sequential Monte Carlo methods, 2010.
[239] C. Evers, J.R. Hopgood, Parametric modelling for single-channel blind dereverberation of speech from a moving speaker, IET Signal Process. 2 (2) (2008) 59–74.
[240] A. Salekin, Z. Chen, M.Y. Ahmed, J. Lach, D. Metz, K. De La Haye, B. Bell, J.A. Stankovic, Distant emotion recognition, Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 1 (3) (2017) 96.
[241] J.B. Alonso, J. Cabrera, C.M. Travieso, K. López-de Ipiña, A. Sánchez-Medina, Continuous tracking of the emotion temperature, Neurocomputing 255 (2017) 17–25.
[242] J.B. Alonso, J. Cabrera, M. Medina, C.M. Travieso, New approach in quantification of emotional intensity from the speech signal: emotional temperature, Expert Syst. Appl. 42 (24) (2015) 9554–9564.
[243] H. Cao, R. Verma, A. Nenkova, Speaker-sensitive emotion recognition via ranking: studies on acted and spontaneous speech, Comput. Speech Lang. 29 (1) (2015) 186–202.
[244] S. Gupta, M.S. Fahad, A. Deepak, et al., Pitch-synchronous single frequency filtering spectrogram for speech emotion recognition, arXiv preprint, arXiv: 1908.03054, 2019.
[245] P. Harár, R. Burget, M.K. Dutta, Speech emotion recognition with deep learning, in: 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN), IEEE, 2017, pp. 137–140.
[246] S. Steidl, A. Batliner, B. Schuller, D. Seppi, The hinterland of emotions: facing the open-microphone challenge, in: 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, IEEE, 2009, pp. 1–8.

Md Shah Fahad (First and Corresponding author) is currently working toward the PhD degree in computer science and engineering at the National Institute of Technology Patna, India. His research interest include machine learning, speech processing, emotions and expressive speech analysis, and pattern recognition.

Ashish Ranjan (2nd author) is currently working toward the PhD degree in computer science and engineering at the National Institute of Technology Patna, India. His research interest include machine learning, Deep learning and computational biology.

Jainath Yadav (3rd author) received the M. Tech and PhD from Indian Institute of Technology Kharagpur. He is an assistant professor with the Department of Computer Science, Central University of South Bihar, India. His research interest include Speech signal processing, emotions and expressive speech analysis, machine learning, deep learning. More information about him can be found at: https://www.cusb.ac.in/index.php/105-faculty-profile/department-of-computer-science/313-mr-jainath-yadav.

Akshay Deepak (4th author and supervisor) received the MS and PhD degrees in computer science from Iowa State University. He is an assistant professor with the Department of Computer Science and Engineering, National Institute of Technology Patna, India. His research interest include data mining, machine learning, emotions and expressive speech analysis and computational biology. More information about him can be found at: http://www.nitp.ac.in/php/faculty_profile.php? id=akshayd@nitp.ac.in