International Journal of Speech Technology
https://doi.org/10.1007/s10772-018-9493-x

Speech emotion recognition research: an analysis of research focus

Mumtaz Begum Mustafa[1] · Mansoor A. M. Yusoof[2,3] · Zuraidah M. Don[4] · Mehdi Malekzadeh[5]

Received: 10 March 2017 / Accepted: 11 January 2018

Abstract

This article analyses research in speech emotion recognition ("SER") from 2006 to 2017 in order to identify the current focus of research, and areas in which research is lacking. The objective is to examine what is being done in this field of research. Searching on selected keywords, we extracted and analysed 260 articles from well-known online databases. The analysis indicates that SER research is an active field of research, dozens of articles being published each year in journals and conference proceedings. The majority of articles concentrate on three critical aspects of SER, namely (1) databases, (2) suitable speech features, and (3) classification techniques to maximize the recognition accuracy of SER systems. Having carried out association analysis of the critical aspects and how they influence the performance of the SER system in term of recognition accuracy, we found that certain combination of databases, speech features and classifiers influence the recognition accuracy of the SER system. We have also suggested aspects of SER that could be taken into consideration in future works based on our review.

Keywords Speech emotion recognition · Emotional speech · ASR system · Emotional speech database · Speech feature · Classification of emotion · Trend analysis

1 Introduction

The recognition of emotion in speech has been the focus of research for some decades, as the inclusion of emotion in human–machine interaction can improve the interaction experiences of users (Picard and Picard 1997). Although humans can express emotion in a multitude of ways, the expression of emotion in speech is considered as one of the most effective (Anagnostopoulos et al. 2012; Batliner et al. 2011; Benzeghiba et al. 2007; Athanaselis et al. 2005; Ten Bosch 2003). The automatic recognition of emotion is a multidisciplinary research area involving different forms of input for recognising emotion (Zeng et al. 2009). In (D'Mello et al. 2012), it was found that most of the research in affective computing focuses on facial expressions (77%) and acoustic-prosodic cues (77%). Almost a third (30%) of the studies tracked body movements, postures, and gestures. Audio-visual features constituted the most common multimodal systems, followed by a trimodal face + voice + posture and gesture. Though there is increasing interest in the recognition of emotion taking a multimodal approach, the use of speech as the primary input to detect emotion in affective computing is more convenient, using existing input devices such as telephone, smartphone and microphones.

✉ Mumtaz Begum Mustafa
mumtaz@um.edu.my

Mansoor A. M. Yusoof
mansoorali@mahsa.edu.my

Zuraidah M. Don
zuraida@um.edu.my

Mehdi Malekzadeh
me.malekzadeh@gmail.com

[1] Department of Software Engineering, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia

[2] Faculty of Business Finance and Hospitality, Mahsa University, Selangor, Malaysia

[3] Department of Operation and Management Information System, Faculty of Business and Accountancy, University Malaya, 50603 Kuala Lumpur, Malaysia

[4] Department of English Language, Faculty of Languages and Linguistics, University of Malaya, 50603 Kuala Lumpur, Malaysia

[5] Department of Software Engineering, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia

Published online: 22 January 2018

Getting the computer to recognize emotions in speech is a challenge, as speech conveys a large amount of information, including explicit linguistic information relating to emotion, and implicit paralinguistic information (Rehmam et al. 2015). Linguistic information includes speech patterns relating to the speaker's words, while paralinguistic information corresponds to variations in the waveform which are independent of words (Anagnostopoulos et al. 2012; Benzeghiba et al. 2007; Calvo and D'Mello 2010). Although research in this field has been underway for more than a decade, prospects for a working system are still far away. This is because of the sheer difficulty of recognizing emotion in speech, a task which is often not easy even for humans.

A typical automatic speech emotion recognition (SER) system recognizes emotion in speech using sound processing alone, without any linguistic information (Altun and Polat 2009). The first daunting task facing researchers is to identify the most appropriate information which can be extracted from speech, and which can be used in the computational identification and discrimination of emotions (Anagnostopoulos et al. 2012; Batliner et al. 2011; Benzeghiba et al. 2007; Athanaselis et al. 2005; Ten Bosch 2003). Secondly, classification techniques are essential to recognize emotions in speech and classify the emotions, and yet there is no consensus among researchers on what the most effective classification techniques might be (Ayadi et al. 2011; Womack and Hansen 1999).

This paper analyses and provides a comprehensive analysis on the research focus carried out within the domain of SER over the period of 2006–2017. Although there have been several articles reviewing existing research in the field, it seeks to contribute to the existing literature by providing analytical information on the focus and outcome of SER research over that period. The knowledge gained from this analysis could be invaluable in gaining insight into the progress of the SER over the last decade. This research also helps to identify the research area within SER that are currently lacking.

A direct comparison of performances of systems developed and reported in the existing research is not viable due to the lack of consistency in the way these methods are designed and assessed in each of the research. However, a generic comparison based on several critical aspects such as database, speech features, and classification of emotion can shed some light on the performance of SER system in recognising emotional speech. This paper also does not provide an exhaustive bibliographical listing of all research papers published over the past decade; what it does is to concentrate on selected key issues of current interest in SER systems development. The aim is to identify the current research focus and its impact on SER.

The paper is organized as follows: Sect. 2 provides an overview of SER research. Section 3 discusses the search process to identify and select suitable published articles. Section 4 presents the findings of this review, and Sect. 5 discusses the major findings and the future research directions. Section 6 concludes the paper.

2 Background

We have identified several existing review papers on SER systems, many of which focus on speech feature and the classification of emotions (Anagnostopoulos et al. 2012; Batliner et al. 2011; Benzeghiba et al. 2007; Ayadi et al. 2011; Ververidis and Kotropoulos 2006). Other key issues discussed include databases of recorded emotional speech and the types of emotion being investigated (Ayadi et al. 2011; Ververidis and Kotropoulos 2006; Cowie et al. 2000).

The major challenges for SER include the definition of emotion and the categorisation of emotions in speech (Anagnostopoulos et al. 2012; Ayadi et al. 2011; Cowie et al. 2000). Human emotion has multiple definitions in different disciplines over a long period of time. Currently, most speech emotion recognition research is based on the work of Ekman (1999), who claimed that there are six basic emotions—happiness, sadness, surprise, fear, anger and disgust—that can be recognized universally. Other research communities focus on the categorization of emotion on the two-dimension valence and arousal (Cowie et al. 2000). The valence dimension refers to the polarity of emotions i.e. how positive or negative they are, and ranges from unpleasant feelings to pleasant feelings of happiness. The arousal dimension refers to the strength of emotions on a scale from apathy to excitement, ranging from sleepiness or boredom to frantic excitement (Nicolaou et al. 2011). Although most existing emotional speech databases are based on Ekman's six basic emotions (Ekman 1999), the two dimensional approach is also widely applied in cross-corpus emotion recognition (Schuller et al. 2010).

In much SER research, the recognition of emotion is made automatic with the development or use of automatic speech recognition (ASR) systems (Altun and Polat 2009; Gharsellaoui et al. 2015). These systems required some form of pre-recorded speech database as the basis for model development. Most research uses pre-recorded databases of acted and predefined speech (Ayadi et al. 2011; Ververidis and Kotropoulos 2006; Cowie et al. 2000) due to the simplicity of the data acquisition. On the other hand, some researchers suggest that acted speech may not represent the genuine expression of emotion in speech (Anagnostopoulos et al. 2012; Ververidis and Kotropoulos 2006; Li et al. 2015). Systems developed using acted databases may lack robustness when recognising emotion in natural speech.

Many reviews and survey papers have identified speech feature as a critical aspect of SER (Anagnostopoulos et al.

2012; Batliner et al. 2011; Benzeghiba et al. 2007; Ayadi et al. 2011; Ververidis and Kotropoulos 2006). Since speech combines complex linguistic and non-linguistic information, the selection and extraction of appropriate speech features is vital for SER systems. Features extracted from speech are of two kinds, low-level-descriptors (LLDs) and functional features (Anagnostopoulos et al. 2012; Tamulevicius and Liogiene 2015). LLDs include prosodic features, and spectral features and their derivatives such as pitch (F0), energy, formants, mel frequency cepstral coefficients (MFCCs) and linear prediction cepstral coefficients (LPCC). Functionals include statistical derivatives such as mean, maximum and minimum, and the zero-crossing rate (Anagnostopoulos et al. 2012).

El Ayadi et al. (2011) on the other hand, state that speech features can be divided into local and global features. Local features refer to speech signals at small intervals called frames, which includes features like pitch and energy. On the other hand, global features are statistical measures of all speech features extracted from an utterance. El Ayadi et al. (2011) also have categorised speech features into four categories, which are continuous features (pitch, energy, formants), qualitative features (voice quality ranging from harsh to tense and breathy), spectral features (LPC, MFCC, LFPC), and TEO (teager energy operator)-based features (TEO-FM-Var, TEO-Auto-Env and TEO-CB-Auto-Env).

Selected features from speech need to be classified by means of pattern recognition techniques such as Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs), support vector machines (SVMs), Artificial Neural Networks (ANNs), decision trees, or k-nearest neighbor distance classifiers (kNNs). Some SER research apply a single classifier while some apply multiple classifiers (Anagnostopoulos et al. 2012; Batliner et al. 2011; Benzeghiba et al. 2007; Ayadi et al. 2011; Ververidis and Kotropoulos 2006; Gharavian et al. 2013; Milton and Selvi 2014; Utane and Nalbalwar 2013).

3 Research aim and approach

The primary aim of this article is to examine current research directions in the domain of SER by focusing on three critical aspects of SER, namely databases, speech features, and classification. Though many of previous projects have highlighted the importance of these three aspects, this research contributes by identifying the actual trend and preferences and how these preferences impact the performance of the automated SER system in term of recognition accuracy. This paper also highlights areas in which research has not been done, in order to indicate possible directions for future research.

There are two research questions to be answered:

- RQ1: What has been the focus of research in SER over the past 12 years in terms of databases, speech features, and classification of emotion?

  To answer this question, we extracted articles relating to SER from 2006 to 2017 inclusive, and made a numerical analysis of the research focus over the 12-year-period.
- RQ2: How have these critical aspects or combination of them contributed to the performance of existing SER systems in terms of recognition accuracy?

  To answer this question, we carried out numerical trend analysis on the extracted articles according to the three critical aspects of SER systems, namely databases, speech features and classification of emotion, and their associations with recognition accuracy using the WEKA toolkit (Brooks et al. 2016).

3.1 Search methodology

The search methodology applied in this research is based on Kitchenham (2004). Specific keywords were used to search for the relevant literature. The search criteria include the process of finding papers on the basis of following keywords:

- Speech emotion recognition
- Automatic speech emotion recognition

The search inclusion criteria applied to filter the initial hits are as follows:

- Publication date: between 2006 and 2017 inclusive
- Search domain: science, technology or computer science
- Publication type: journals, proceedings and transactions
- Article type: full text and reviews
- Subject: directly addresses one or more parts of the analysis framework
- Language: English

Exclusion criteria:

- Studies that do not focus explicitly on the machine recognition of emotion in speech
- Studies that discuss recognition of emotion in speech as a side topic
- Studies on emotional speech without actual recognition
- Studies that review the work of others
- Studies that do not provide details of their experiments or experimental design
- Opinions, viewpoints, keynotes, discussions, editorials, comments, tutorials, prefaces, and anecdotal papers and presentations in slide format without any associated papers

Table 1 The distributions according to databases and the number of papers

Digital database libraries	Keywords and hits	
	Speech emotion recognition	Automatic speech emotion recognition
Science direct	9	14
IEEE explore digital library	40	84
Springer link	7	23
ISI web of knowledge	11	38
Google scholars	9	25

N = 260

4 The findings of the review

A total of 260 papers were extracted using the key words. Of these, 157 were conference papers, and the remaining 103 journal articles. Table 1 shows the distributions according to publication type and the number of papers. The largest numbers of papers (47%) were obtained from IEEE explore digital library. Figure 1 shows the breakdown of the papers extracted for each of the 12 years. The highest number was for 2012, and the lowest for 2014. The reduced number of papers for 2016 and 2017 could be attributed to the shift in research focus towards the audio–visual recognition of emotion.

4.1 RQ1: what are the research focus on SER over the past 12 years in term of the critical aspects namely, databases, speech features, and classification?

The 260 papers were divided according to their focus into several categories, including classification, noise, database, features, adaptation, variability, and speaker independent recognition. Figure 2 shows the research focus of the papers extracted each year, and indicates that the focus has been on databases (11%) speech feature selection and extraction (42%), and classification (34%). In total, these three aspects accounted for more than 87% of the extracted articles. Other issues raised are speaker independence (7%) and adaptation which includes cross-corpus adaptation, cross-lingual adaptation, and neutral speech to emotional speech adaptation (3%). Issues rarely considered include noise; performing SER in noisy environment especially for real time speech emotion recognition (1.5%) and dealing with the variability in emotional speech particularly for the cross-corpus SER research (1.0%). A total of 315 issues were identified in the 260 papers, indicating that some of the papers focused on more than one issue.

4.1.1 Databases

One of the major issues of debate in SER research is the selection of databases for the training and testing of automatic SER systems. It is often claimed that the best kind of database, particularly for evaluating the performance of SER systems, consists of spontaneous speech collected in real life situations (Anagnostopoulos et al. 2012; Cowie et al. 2000; Li et al. 2015; Douglas-Cowie et al. 2003, 2007; Kostoulas et al. 2007; Navas et al. 2006; Batliner et al. 2006; Devillers and Vidrascu 2006; Kim et al. 2006; Zhang and Zhao 2008; Steidl et al. 2008; Neiberg and Elenius 2008; Barra Chicote et al. 2009; Erdem et al. 2010; Espinosa et al. 2010; Sztahó et al. 2011; Tarasov and Delany 2011; Bozkurt et al. 2011; Polzehl et al. 2011; Attabi and Dumouchel 2012, 2013; Atassi et al. 2012; Deng et al. 2012, 2014; Planet and Iriondo 2012; Feraru and Zbancioc 2013; Sethu et al. 2013; Busso et al. 2013; Alam et al. 2013; Le and Provost 2013;

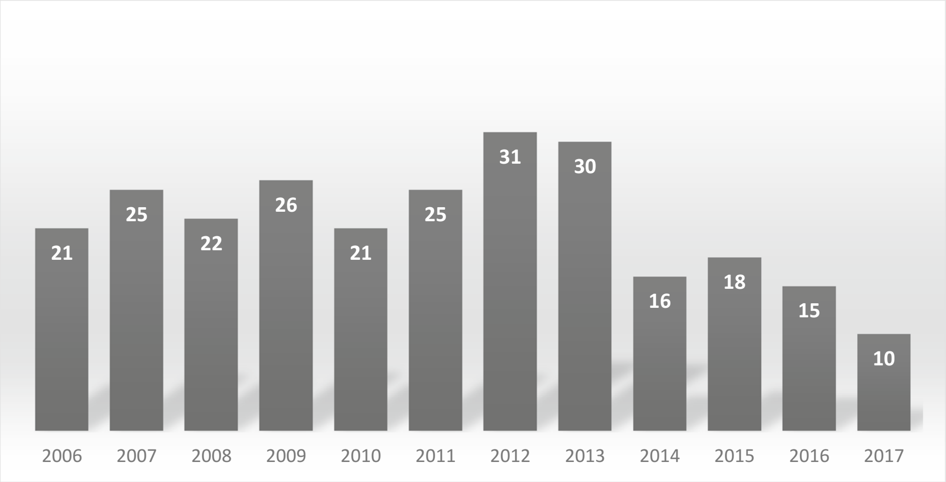

Fig. 1 The breakdown of papers extracted in each of the 12 years

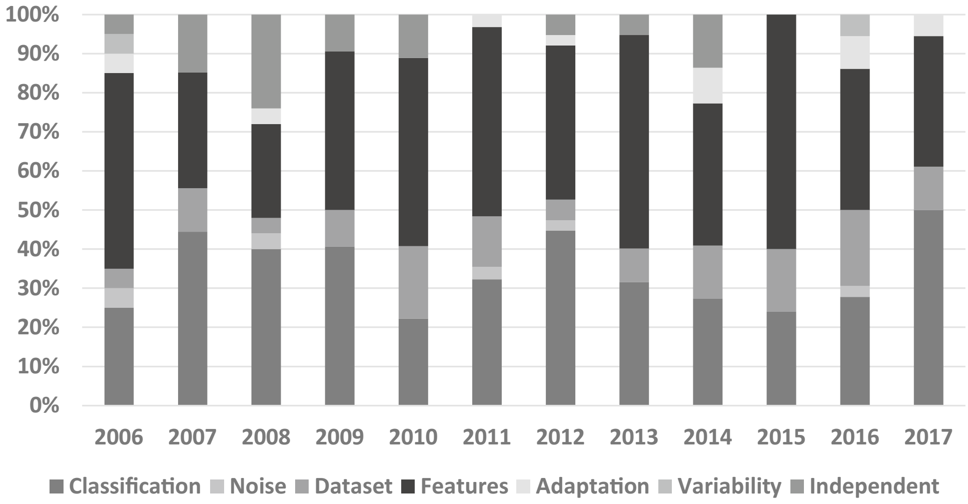

Fig. 2 The analysis of research focus of the papers extracted from 2006 to 2017

Zheng et al. 2015; Trigeorgis et al. 2016; Schmitt et al. 2016; Pohjalainen et al. 2016; Zha et al. 2016; Cummins et al. 2017; Mencattini et al. 2017). However, about 78% of the papers in the present study report the use of acted emotional speech (Han et al. 2014, 2017; Xiao et al. 2006, 2007, 2009; Zhou et al. 2006, 2009; Huang and Ma 2006; You et al. 2006; Luengo et al. 2010; Vogt and André 2006, 2009; Hu et al. 2007; Liu et al. 2007; Soltani and Ainon 2007; Pao et al. 2007, 2012; Lugger and Yang 2007; Gamage et al. 2017; Abdelwahab and Busso 2017; Le et al. 2017; Iriondo et al. 2007; Sedaaghi et al. 2007; Álvarez et al. 2007; Batliner et al. 2007; Mao et al. 2007, 2009, 2016; Sethu et al. 2007, 2008a, b; Truong and Leeuwen 2007; Zhu et al. 2017; Hussain et al. 2017; Schuller et al. 2007; Ye et al. 2008; Fu et al. 2008a, b; Morales-Perez et al. 2008; Deng et al. 2013, 2017; Atassi and Esposito 2008; Iliev and Scordilis 2008; Mannepalli et al. 2016; Bertero and Fung 2017; Schuller 2008; Albornoz et al. 2008, 2011; Scherer et al. 2008; Ringeval and Chetouani 2008; Ser et al. 2008; Dai et al. 2008; Casale et al. 2008, 2010; Huang et al. 2014, 2016; Lugger et al. 2009; Wenjing et al. 2009; Vondra and Vích 2009; Chenchah and Lachiri 2014; Chandrakala and Sekhar 2009; Sun et al. 2009, 2015; Mishra and Sekhar 2009; Hassan and Damper 2009; Shah 2009; Kim et al. 2009, 2011, 2012; Iliou and Anagnostopoulos 2009, 2010a, b; Schwenker et al. 2009; Chandaka et al. 2009; Chakraborty et al. 2016; Song et al. 2016; Fayek et al. 2016; Shaw et al. 2016; Lim et al. 2016; Zhang et al. 2010; Kotti et al. 2010; Böck et al. 2010; Gharavian et al. 2010, 2012; Weninger et al. 2016; Bozkurt et al. 2010; Koolagudi et al. 2010; Iliev et al. 2010; Brester et al. 2016; Chavhan et al. 2010, 2015; Sagha et al. 2016; Khanna and Kumar 2011; Ananthakrishnan et al. 2011; Koolagudi and Krothapalli 2011; Pathak and Kulkarni 2011; Wang et al. 2011; Zong et al. 2016; He et al. 2011; Shen et al. 2011; Yeh et al. 2011; Glüge et al. 2011; Shaukat and Chen 2011; Swain et al. 2015; Jin et al. 2015; Jeon et al. 2011; Lee and Tashev 2015; Busso et al. 2011; Khan et al. 2011; Sun and Wen 2015; Sidorov et al. 2014; Philippou-Hübner et al. 2012; Yüncü et al. 2014; Zbancioc and Feraru 2012; Christina and Milton 2012;Chen et al. 2012; Bojanić et al. 2012; Arias et al. 2013, 2014; Milton and Selvi 2014; Sheikhan et al. 2012, 2013; Ntalampiras and Fakotakis 2012; Jiang et al. 2012; Henríquez et al. 2014; Kamińska and Pelikant 2012; Hamidi and Mansoorizade 2012; Yang et al. 2012; Elbarougy and Akagi 2012, 2013; Giannoulis and Potamianos 2012; Thapliyal and Amoli 2012; Georgogiannis and Digalakis 2012; Yun and Yoo 2012; Cheng and Duan 2012; Ivanov and Riccardi 2012; Delic et al. 2012; Pan et al. 2012; Balti and Elmaghraby 2013; Přibil and Přibilová 2013; Seehapoch and Wongthanavasu 2013; Javidi and Roshan 2013; Shah et al. 2013; Garg et al. 2013; Bhaykar et al. 2013; Li et al. 2013; Kishore and Satish 2013; Rao et al. 2013; Chiou and Chen 2013; Milton et al. 2013), while some researchers use induced speech database (Koolagudi and Krothapalli 2012; Wu et al. 2006). In some SER research, researches make use of more than one types of speech database (Xiao et al. (2006); Shami and Verhelst (2007); Grimm et al. (2007a, b); Wagner et al. (2007); Morrison et al. (2007); Kandali et al. (2008); Scherer et al. (2008); Busso et al. (2009); Bozkurt et al. (2009); Rong et al. (2009); Wu et al. (2009); Schuller et al. (2010); Vogt and André (2011); Fernandez and Picard (2011); Vlasenko et al. (2011a, b); Wu et al. (2011); Koolagudi and Rao (2012); Rao et al. 2012).

Acted speech is used in view of its many advantages. First, recordings are made in a noise-free environment, which makes feature extraction much simpler than in the case of naturally produced data. Secondly, acted databases are more controlled and have been labelled for certain types of emotional expressions. Labelling an emotional speech database is very important for classification and determining the performance of classifiers in recognising emotional

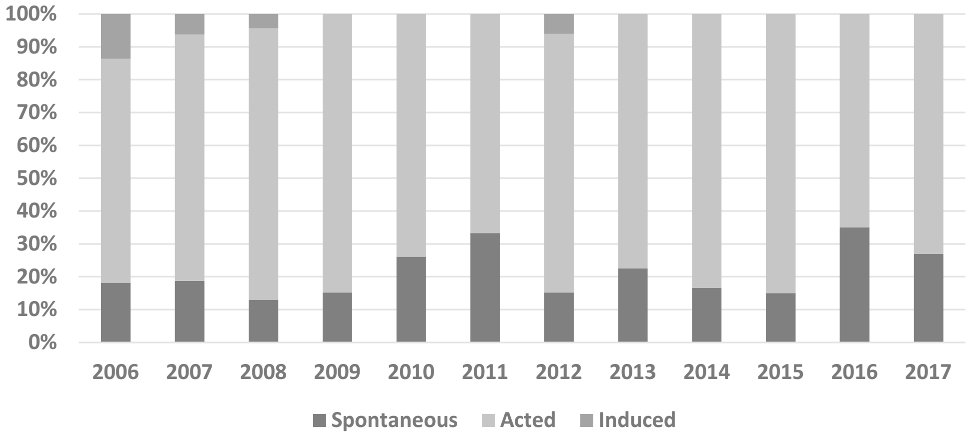

Fig. 3 The analysis of the databases of the papers extracted from 2006 to 2017

Table 2 The language of the emotional speech databases used in the SER research from 2006 to 2017

Language	Year												
	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	Average
German (%)	22	48	45	53	63	60	38	49	41	36	46	31	44
English (%)	6	6	27	22	7	23	9	9	30	18	19	23	17
Spanish (%)	6	6	5	3	0	0	0	0	4	0	0	0	2
Danish (%)	6	3	0	3	7	0	0	3	4	0	0	0	2
Dutch (%)	0	6	0	0	0	3	0	0	0	5	0	0	1
Other European (%)	11	3	0	0	0	3	15	11	11	9	27	23	9
Mandarin (%)	39	23	18	13	7	3	15	6	4	9	0	8	12
Other Asian (%)	11	3	5	6	11	7	24	23	7	23	8	15	12
African (%)	0	0	0	0	4	0	0	0	0	0	0	0	0

speech (Litman and Forbes-Riley 2006). Acted speech has the advantage for labelling and annotation, as text for recording are available before the recordings are made, so that labelling is much easier (Douglas-Cowie et al. 2003, 2007). By contrast, spontaneous databases are normally created from recordings of actual human interaction and so contain extraneous noises (Douglas-Cowie et al. 2003, 2007; Kostoulas et al. 2007; Navas et al. 2006), which make it much more difficult to do the labelling. In (Schuller et al. 2007), the labelling was done by several labellers with expert knowledge in linguistics, and majority voting was used to decide on the labels for particular words. The combine efforts of multiple labellers can increase the accuracy of the labels for SER research. Recently, several of the SER researchers choose an additional set of unlabelled testing samples from target speech corpus and use them to serve as an auxiliary set to train the labelled training samples from source speech corpus (Song et al. 2016; Zong et al. 2016). On top of that, the notion of cross-lingual SER is also gaining attraction among researchers, where the training and testing emotional speech database are based on two different languages (Sagha et al. 2016).

Figure 3 lists the databases used in research between 2006 and 2017. The use of spontaneous databases began to increase in 2010 and 2011 but fell again from 2013 to 2015. This decrease could be attributed to new kinds of SER research on languages not previously attempted, as indicated in Table 2. Spontaneous speech in SER gains new attention in the year 2016 and 2017. However, we found that induced speech database was no longer the focus in SER research 2013.

76% of the databases used for SER over the last 12 years are from Europe, with more than 44% using German emotional databases, followed by English (17%) emotional speech databases. The German databases (Anagnostopoulos et al. 2012) were chosen because they had been adequately labelled for speech feature extraction and classification. In addition to Mandarin (12%), SER research has been carried out on other Asian languages including Farsi, Telegu, Japanese, Hindi, and Korean (Erdem et al. 2010; Espinosa et al. 2010; Sun et al. 2009; Wang et al. 2011; Bhaykar et al. 2013). Table 2 shows the decline in the use of German databases in recent years as more emotional speech databases have become available for other European and Asian languages. Our analysis also shows that RECOLA, a

Table 3 The speech feature categories applied in the existing SER research from 2006 to 2017

Speech features categories	Year												
	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	Average
Continuous (%)	11	35	23	15	10	17	13	17	13	10	9	11	15
Spectral (%)	33	35	23	35	33	25	19	33	25	30	36	32	30
Qualitative (%)	0	4	5	0	0	8	0	0	0	0	0	0	1
TEO (%)	0	0	0	0	0	0	0	0	0	0	0	0	0
Continuous and spectral (%)	39	15	27	31	33	42	55	30	44	35	47	49	37
Continuous and qualitative (%)	0	4	5	0	5	0	0	0	0	5	0	0	2
Spectral and qualitative (%)	6	0	0	0	0	0	0	3	0	0	0	0	1
TEO and spectral (%)	0	0	0	0	0	0	3	0	0	0	0	0	0
Continuous, spectral and qualitative (%)	11	4	5	0	0	8	10	7	6	10	8	8	6

French SER database is gaining popularity among the SER researchers particularly for classifying emotion according to valence and arousal (Trigeorgis et al. 2016; Schmitt et al. 2016; Pohjalainen et al. 2016; Cummins et al. 2017; Mencattini et al. 2017; Han et al. 2017; Le et al. 2017; Weninger et al. 2016). It was also observed that very little SER research has been done on African languages, only one project being reported in 2010.

4.1.2 Speech feature

The recognition of emotion in speech by machine requires the identification and extraction of speech features that can be used to discriminate and classify a particular emotion. The selection of the right features is critical in determining the effectiveness of SER systems (Anagnostopoulos et al. 2012; Batliner et al. 2011; Benzeghiba et al. 2007; Altun and Polat 2009; Ayadi et al. 2011; Ververidis and Kotropoulos 2006; Tamulevicius and Liogiene 2015; Esmaileyan and Marvi 2014; Rodríguez et al. 2013; Tahon et al. 2015; Wu and Liang 2011; Yang and Lugger 2010). Researchers have long struggled to balance the number of speech features and the computational cost of feature extraction and classification (Anagnostopoulos et al. 2012; Ayadi et al. 2011; Ververidis and Kotropoulos 2006).

The use of more features typically improves the accuracy of the SER system (Anagnostopoulos et al. 2012; Batliner et al. 2011; Benzeghiba et al. 2007; Altun and Polat 2009; Ayadi et al. 2011; Ververidis and Kotropoulos 2006; Tamulevicius and Liogiene 2015; Esmaileyan and Marvi 2014; Rodríguez et al. 2013; Tahon et al. 2015; Wu and Liang 2011; Yang and Lugger 2010). However, too many features can slow down the recognition process in view of the more complex operations required to process them. Some of the more common features used to discriminate emotion in speech were LLDs (Anagnostopoulos et al. 2012; Wu and Liang 2011; Yang and Lugger 2010; Firoz Shah et al. 2009; Kostoulas et al. 2010; Vlasenko et al. 2007) such as Mel-frequency cepstral coefficients (MFCC), pitch (F0), Energy (intensity). The 260 articles yielded a total of 594 feature types, which means that each project used two or three types on average, which may indicate the optimum level for SER. From the work of El Ayadi (2011), we have analysed the speech features used in existing SER research into four categories which are continuous features (pitch-related features, formants features, energy-related features, timing features, and articulation features), qualitative features (voice quality features, harsh, tense, and breathy), spectral features (LPC, MFCC, and LFPC), and Teager-energy-operator (TEO)-based features (TEO-decomposed FM variation, normalized TEO autocorrelation envelope area, and critical band-based TEO autocorrelation envelope area) as presented in Table 3. The majority of the SER focuses on the LLDs including temporal descriptors (such as energy, zero crossing rate), spectral descriptors (spectral centroid, spectral width, spectral asymmetry, and spectral flatness), cepstral descriptors (mel-frequency cepstral coefficient), perceptual descriptors (loudness), and so on.

Researchers have recently applied appropriate techniques to select the best speech features to recognise a particular emotion such as GMM and SVM (Gaurav 2008) which enables the SER to maximise the recognition of emotion in speech based on specific pre-conditions set by the researchers.

In order to identify emotions, features of speech must be associated with particular emotions. The most common approach to emotion classes is based on the work of Plutchik (1991), who proposed a set of eight basic emotions, anger, sadness, happiness, fear, disgust, trust, anticipation, and surprise. Plutchik's work is also the source of the two-dimensional classification of emotions according to valence and arousal, and the plotting of emotions on the wheel of emotion (Plutchik 1991). The emotions that are regularly investigated in SER are presented in Table 4.

Table 4 Types of emotion in the existing SER research from 2006 to 2017

Emotion	Year												
	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	Average
Anger (%)	21	18	18	17	17	16	17	16	16	14	14	19	17
Anxiety (%)	0	2	1	1	3	1	0	2	0	4	11	7	3
Boredom (%)	6	9	10	8	10	9	5	7	5	4	11	7	8
Disgust (%)	5	8	6	8	5	6	8	10	12	8	11	7	8
Fear (%)	12	11	7	9	8	9	10	10	12	9	11	14	10
Happiness (%)	12	11	17	11	11	9	12	10	11	11	14	19	12
Joy (%)	3	7	1	3	4	4	4	3	3	3	0	0	3
Neutral (%)	9	8	16	15	14	14	17	15	14	13	14	12	13
Sadness (%)	17	13	16	15	11	13	14	12	13	12	14	14	14
Surprise (%)	6	2	4	2	3	3	4	3	8	6	0	0	3
Others negative (%)	6	5	4	2	5	6	4	4	2	8	0	0	4
Others positive (%)	2	6	2	7	8	9	5	9	2	7	0	0	5

A total of 1460 emotion tokens were investigated in the 260 papers, which means that on average five to six types were to be recognized in each case. Anger was the most commonly evaluated emotion, possibly because it has the highest correlation with the features regularly used in SER research. Other emotions commonly investigated include neutral, sadness and happiness. Joy and happiness are wonderful feelings to experience but are very different. Joy is more consistent and is cultivated internally. It comes when you make peace with yourself, why you are and how you are, whereas happiness tends to be externally triggered and is based on other people, things, places, thoughts, and events (Tseng et al. 2005). It is possible that some researchers use the term joy instead of happiness as wonderful feelings to experience from within and not from external stimuli. However, no SER researches that attempt to classy emotion as happiness and joy in the same experiment.

Other positive emotions (empathy, motherese, amusement, and satisfaction) and other negative emotions (irritation, disappointment, stress, and embarrassment) have been evaluated in some SER researches. Researchers have recently paid attention to other types of emotion not previously investigated, such as excitement, frustration and emphasis. On top of the basic emotion as suggested by Plutchik (1991), SER researches also categorised emotion as valence and arousal (Trigeorgis et al. 2016; Schmitt et al. 2016; Pohjalainen et al. 2016; Mencattini et al. 2017; Han et al. 2017; Le et al. 2017). Our research shows that SER researches in the last 5 years make use of standard database such as the Berlin database (Mannepalli et al. 2016; Huang et al. 2016; Lim et al. 2016; Brester et al. 2016; Song et al. 2016; Mao et al. 2016; Zong et al. 2016), focusing on six basic emotions of anger, anxiety, boredom, disgust, fear, and sadness (the Berlin speech database also include neutral speech used in SER research).

4.1.3 Classification of emotion

SER systems rely on machine learning techniques to discriminate and classify emotion in speech by making the best use of the speech features extracted. Classification is equally important to speech feature selection and extraction (Anagnostopoulos et al. 2012; Batliner et al. 2011; Ayadi et al. 2011; Ververidis and Kotropoulos 2006; Bitouk et al. 2009; Gaurav 2008; Sethu et al. 2009; Tabatabaei et al. 2007). Many different classifiers have been proposed to classify emotion in speech; but it is first necessary to consider the types of emotion at the focus of SER research.

Table 5 presents the classifiers used to discriminate emotions based on the selected features, where in some cases, multiple classification techniques were used (Anagnostopoulos et al. 2012; Lefter et al. 2010; Scherer et al. 2008, 2009; Xiao et al. 2007). The most frequently used machine learning technique is the support vector machine (SVM) used in 27% of cases (includes SMV and its variants: support vector regression). SVM is a powerful technique which has the ability to discriminate emotions in speech by making use of the common speech features currently applied in SER research. Other classifiers found to be effective in SER include the mixture model based learners (Gaussian mixture model) at 17% and neural networks (including its variants such as deep neural network, convolutional neural network, radial basis function, recurrent neural network, multilayer perceptron, and recursive neural network) at 12%.

Though SVM, GMM and NN are common in SER, researchers have proposed multi-classifiers which involve the application of more than one classifier to improve the recognition of emotion in speech (Anagnostopoulos et al. 2012; Lefter et al. 2010; Scherer et al. 2008, 2009; Xiao et al. 2007). Multi-classifiers were proposed to take best advantage of each classifier and overcome the disadvantages.

Table 5 Emotion classification techniques in the existing SER research from 2006 to 2017

Classification technique	Year												
	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	Average
Neural network (%)	13	9	18	14	12	16	17	15	28	11	30	41	19
Support vector machine (%)	21	26	26	23	29	30	25	29	29	33	29	25	27
Mixture model based learners (%)	15	15	7	26	18	18	15	23	6	19	0	9	1
Linear regression (%)	3	0	0	0	0	0	0	0	0	0	4	0	1
Naïve bayes (%)	9	9	4	3	3	5	5	0	0	4	4	6	4
Decision tree based learners (%)	12	13	8	13	3	3	0	6	6	11	0	0	6
Hidden markov model (%)	15	11	11	13	11	15	10	13	3	7	4	0	9
K-nearest neighbor (%)	3	11	15	5	8	3	13	6	11	4	0	6	7
Others (%)	9	6	11	3	16	10	15	8	17	11	29	13	12

Table 6 The categorization of the critical aspects for association analysis

Databases	Language	Speech features	Classification	Recognition accuracy
Acted	European	Continuous	Single classifier	High (66.7–100%)
Induced	Asian	Spectral	Multiple classifiers	Medium (33.4–66.6%)
Spontaneous	Other	Qualitative		Low (0–33.3%)
		TEO		
		Continuous and spectral		
		Continuous and qualitative		
		Spectral and qualitative		
		TEO and spectral		
		Continuous, spectral and qualitative		

SER research using multi-classifiers rose from 2007 to 2011 before slowing down in 2012. This is possibly due to the fact that many of the single classifiers can perform equally well, so that the use of more than one classifier does not necessarily increase the recognition accuracy of the SER systems (Anagnostopoulos et al. 2012).

Other classification techniques such as sequential minimal optimization (SMO) and Fisher Kernel have recently been used. Neural Network variants including artificial neural network (ANN) (Chakraborty et al. 2016; Shaw et al. 2016), deep neural network (DNN) (Cummins et al. 2017; Sánchez-Gutiérrez et al. 2014; Le et al. 2017), convolutional neural network (CNN) (Trigeorgis et al. 2016; Bertero and Fung 2017; Fayek et al. 2016; Lim et al. 2016), as well as recurrent neural network (RNN) (Weninger et al. 2016; Glüge et al. 2011; Lee and Tashev 2015), are increasingly applied in SER research. The increased use of alternative classifiers may indicate a search for more robust classifiers that work well with existing speech features.

4.2 RQ2: how these critical aspects or their combinations have contributed to the recognition accuracy of existing SER systems?

SER is a unique research domain in which different projects cannot be compared directly. What is possible is to examine how the critical aspects or their combinations contribute to the performance of SER systems. In SER research, two of the most important performance measures are classification accuracy and recognition accuracy. The focus here is on recognition accuracy, as it reflects the expected performance of SER systems (Athanaselis et al. 2005).

We used the association function in WEKA toolkit version 3.6.31[1] to determine how combinations of the critical factors identified earlier influence the recognition accuracy of the SER system. We classed the recognition accuracy of SER systems as high, medium or low. In making the association analysis, we categorized the identified factors as presented in Table 6.

For the association analysis, we used the Apriori algorithm, and selected a minimum confidence level of 0.9. Table 7 shows the top ten association rules of the identified factors and the recognition accuracy of the SER systems.

[1] http://www.cs.waikato.ac.nz/ml/weka/.

Table 7 Top ten association rules of the identified factors and the recognition accuracy of the SER systems

No	Datasets	Language	Speech features	Classification	Recognition accuracy	Confidence level
1	Acted	–	Spectral	–	High	1.00
2	Acted	–	Spectral	Single	High	1.00
3	Acted	European	–	Single	High	0.99
4	Acted	European	Continuous and spectral	Single	High	0.98
5	–	European	Continuous and spectral	Single	Medium	0.97
6	Acted	–	Continuous and spectral	Single	High	0.94
7	Acted	Asian	Continuous and spectral	Single	High	0.92
8	Acted	–	–	Single	High	0.92
9	–	–	Continuous and spectral	Single	Medium	0.91
10	Acted	European	–	–	High	0.91

5 Discussion

This discussion evaluates SER research from within the research paradigm itself, focusing mainly on the three critical aspects (databases, speech features and classification) of SER mentioned earlier, and considers what has worked, and what has worked less well, and where we go from here.

SER research is a challenging and exciting field in affective computing. This paper has examined the focus in SER research and considered what might be lacking. It was observed that it might not be possible to make direct comparisons in view of variation in experimental design regarding speech feature selection, speech databases, and the classification techniques. Different projects proclaim the merits of their own experiments using measures such as recognition accuracy. This research proposes a novel approach for assessing the influence of the critical aspects (experimental design) and recognition accuracy by categorizing them into common groups as shown in Table 6, and generates association rules that relate the identified factors using a suitable toolkit.

5.1 Databases

From the association analysis, among the top ten association rules, acted speech databases was found to be associated with high recognition accuracy in eight instances. Judging by the data collected from the last 12 years, most published research makes use of acted speech databases. Acted speech databases contribute toward high recognition accuracy as they contain the necessary information to classify an emotion according to selected speech features, and can be labelled more objectively than other databases such as spontaneous databases.

However, the recognition rates reported for acted speech databases may not reflect the effectiveness of the system in real life. Spontaneous speech databases are more reflective of real life examples and measure the true performance of SER systems. However, they suffer from classification problems, and some recorded speech samples are not classified for emotion. The top ten association rules did not show any relationship between the spontaneous speech databases and recognition accuracy. This is probably due to very limited research in SER that make use of spontaneous emotional speech databases. One of the critical issues that need to be addressed in spontaneous speech databases is the labelling of the databases (Navas et al. 2006).

SER researchers are certainly aware of the difference between acted and natural speech, but acted speech is nevertheless routinely used for data out of practical necessity. Although some highly accomplished actors may be able to simulate some emotions realistically on stage, there is no guarantee that actors are able to produce realistic simulations in the artificial conditions of the recording studio. Unless they are able to simulate the physiological responses too, they have to produce some performance that is regarded as an acceptable substitute for the expression of the genuine emotion. Throughout our analysis, we found that the SER research is being more standardized with the use of similar speech databases and speech features. We also noticed that sspontaneous speech in SER research gains new attention in the year 2016 and 2017.

The existing SER research emphasises on the offline SER, using pre-recorded emotional speech databases (acted or spontaneous). Few SER research focused on the real-time speech emotion recognition (Bahreini et al. 2016), as there are many issues that need to be resolved such as noisy environment (Pohjalainen et al. 2016), classes of emotion (Schuller 2008), and extracting continuous speech stream (Kim et al. 2007) among others.

Many of the existing SER work focus on very limited emotion class based on the work of Ekman (1957, 1972, 1999), six emotions with the addition of Plutchik's boredom (which for Plutchik is a mild form of disgust) and anxiety (which belongs to other approaches to basic emotions). However real-time SER systems need to deal with wider range of emotion. For example in (Huang and Ma 2006), fifteen emotions; anxiety, boredom, cold anger, contempt, despair, disgust, elation, happy, hot anger, interest, neutral, panic, pride, sadness, and shame are recognized by the real-time SER system. In (Pathak and Kulkarni 2011; Sun and Wen 2015), adaptation approach was adopted for developing a real-time SER system to overcome the problem of speaker independent (SI) models that are generally used in speech recognition tasks (Sun and Wen 2015).

A real-time SER system needs to overcome the issue of noise. However, many of the speech databases (acted or spontaneous) were recorded in low noise environment. In (Schuller 2008), white noise was added to the samples of recorded emotional speech database, which is common practice in general speech processing tasks, especially in speech and speaker recognition.

In term of languages, there are four instances in which European emotional speech databases are associated with high recognition accuracy as opposed to Asian or other languages. SER research was found to concentrate on specific languages that have good emotional speech databases, and many are available for European languages such as German, English and Spanish. An increasing number of emotional speech databases are becoming available for Asian languages such as Malay (Esmaileyan and Marvi 2014; Harimi et al. 2016) Mandarin, Farsi, Telegu, Japanese, Hindi, and Korean (Erdem et al. 2010; Espinosa et al. 2010; Sun et al. 2009; Wang et al. 2011; Bhaykar et al. 2013). However, very little is available for African and South American languages, with the result that very little SER research has been done for these languages. SER research is likely to continue to concentrate on major European and Asian languages. It will be interesting to examine SER research on third-world languages, despite the lack of essential resources such as emotional speech databases.

The present SER research concentrates on a narrow range of emotions, primarily based on the work of Plutchik (Attabi and Dumouchel 2012), with the focus on a set of emotions including anger, sadness, happiness, fear, disgust, trust, anticipation, and surprise. Though several researchers have suggested taking the two-dimensional approach including valence and arousal, few emotions have actually been plotted on these dimensions. Anger emerged as the most commonly evaluated emotion in SER, possibly because it has the highest correlation with the features that are regularly used in SER research. This trend is likely to continue as the existing emotional speech databases are limited to those eight emotions. It will be interesting for researchers to consider other types of emotion on the arousal and valence dimensions.

A pick-'n'-mix approach to emotion types guarantees that different research findings will not be comparable, and is of doubtful scientific validity. This could explain, at least in part, the lack of progress in the years 2006–17. Using Plutchik's wheel as a starting point would make possible a more systematic approach to the recognition of emotion, to find out, for example, whether paired emotions are ever confused, whether the same phonetic speech features are associated with an emotion at different levels of arousal, or whether the ability to distinguish emotions is related to a measure of the distance between them.

5.2 Speech features

Current SER research seems to be based on the assumption that an emotion of a certain type arises in the mind and has a predictable effect on the speaking voice. If this is so, there must be a set of features that relate emotions to measurable acoustic properties of the waveform. If these features can be discovered, computers can be programmed to recognise emotions automatically in speech. The question is whether this assumption is justified. It is possible to achieve this goal only if there are some reliable acoustic correlates of emotion/affect in the acoustic characteristics of the signal (Pierre-Yves 2003).

Table 3 lists the speech features used in SER research, but it is not clear on what grounds they have been selected. Features described e.g. as "F0" (part of continuous features) are also vague. F0 obviously has some connection with emotion, but it could involve range, the mean, or the rate of change. The rate of change is itself not a simple variable, because the change could be aligned with vowels or longer stretches, and must be shown to differ in some way from normal intonation patterns.

The association analysis found that spectral features (two instances) as well as continuous and spectral speech features (three instances) are associated with high recognition accuracy. Features such as MFCC, pitch, duration and energy enable the SER to recognise emotions with high accuracy as they contain information that can be used by the classifiers to discriminate particular types of emotion. The use of qualitative features was found to increase in recent years, but not much work so far has focused on TEO-based features.

The review shows that LLDs and functional are becoming the standard features for SER research (Le and Provost 2013; Deng et al. 2013, 2014, 2017; Zha et al. 2016; Gamage et al. 2017; Abdelwahab and Busso 2017; Bertero and Fung 2017; Chakraborty et al. 2016; Song et al. 2016; Brester et al. 2016; Mao et al. 2016; Sagha et al. 2016; Zong et al. 2016; Sun et al. 2015; Sun and Wen 2015; Garg et al. 2013; Busso et al. 2009). In fact many of the SER research have applied the standard features made available in the Interspeech Challenges on Emotion in 2009 (Erdem et al. 2010; Tarasov and Delany 2011; Bozkurt et al. 2011; Deng et al. 2012, 2013, 2014; Planet and Iriondo 2012; Le and Provost

2013; Sagha et al. 2016; Zong et al. 2016; Sun et al. 2015; Sun and Wen 2015; Garg et al. 2013; Busso et al. 2009), which includes both continuous and spectral features. These speech features have also been applied as benchmark in several SER research (Sethu et al. 2013; Attabi and Dumouchel 2013; Deng et al. 2014, 2017; Song et al. 2016; Brester et al. 2016; Mao et al. 2016). As the SER community adopted standard feature for recognizing emotion, comparison can be made on the techniques that can improve the recognition accuracy of the SER systems.

For the period 2016–2017, despite the popularity of LLDs, new forms of features are being considered in SER such as Coiflet Wavelet Packet Cepstral Coefficients (CWPCC) (Lim et al. 2016), and bag of audio word (BoAW) (Schmitt et al. 2016; Gamage et al. 2017). However, their application in SER is limited, and little comparison on their performance was made with the LLDs. From the analysis of the SER research in the past 12 years, it is highly likely that the focus will be on the LLDs, and at the same time any new features to be proposed by researchers will likely be benchmarked with the LLDs.

5.3 Classification of emotion

In SER research, the selection and use of suitable classifiers is essential for recognition accuracy. Although many different techniques have been proposed, the most frequent are HMM, SVM, GMM and NN. These are very powerful machine learning techniques and their effectiveness has been proved in existing research. Our analysis indicates that the use of a single classifier enabled high recognition accuracy in six instances. This indicates that a single classifier may be sufficient to discriminate emotions in speech, which is important because the use of many classifiers is computationally expensive. It is also to be observed that there has been a recent increase in the use of new types of classification techniques such as convolutional neural network (CNN) (Trigeorgis et al. 2016; Bertero and Fung 2017; Fayek et al. 2016; Lim et al. 2016), recurrent neural network (RNN) (Weninger et al. 2016; Glüge et al. 2011; Lee and Tashev 2015), and deep believe (Le and Provost 2013; Mannepalli et al. 2016; Huang et al. 2016). It is highly likely that the SER researchers will continue to use these classification techniques in the near future. Most of these classifiers are the extension of neural networks that have the ability to handle complex tasks such as vision, language, and other AI-level tasks (Feraru and Zbancioc 2013), and is used to recognise emotion both in speech alone and in audio-visual data. Researchers are constantly looking for more robust classifiers that can work well with the existing speech features such as continuous and spectral features. At the same time, some of the more traditional techniques such as the HMM and GMM are becoming more irrelevant in the SER research.

The association analysis indicates that a single classifier combined with acted emotional speech and spectral features leads to high recognition accuracy with a confidence level of 1.00. For European language speech databases, the combination of acted speech and single classifier together with continuous and spectral features contribute to the high recognition accuracy of emotion from speech (confidence level 0.99–0.98). For Asian language speech databases, the combination of acted speech, a single classifier and continuous and spectral features contributes to the high recognition accuracy of emotion from speech (confidence level 0.92).

5.4 The future direction

In view of the large amount of time and effort devoted to SER research, one would expect an overall improvement in results. However, this was not always the case. In order to understand why this should be, it is necessary to make an external evaluation of work in the field. From a wider scientific perspective, the review of SER research raises a number of issues including the relationship between genuine and simulated emotion, and the nature of emotions.

The problem is that what counts as an acceptable substitute depends on culture, and is subject to cultural change, so that what is accepted at one time as a realistic or iconic representation of the expression of emotion might be regarded within a few decades as not realistic or representative at all. This raises the serious question whether SER research is investigating the expression of genuine emotion or the expression of currently accepted substitutes.

The use of standard database such as the Berlin database improves the comparability of SER research particularly on the effectiveness of proposed approach. With the advancement in the methods in SER for several well-resourced languages, other languages did not share the same advancement. The lack of resources is always the major issue in SER. Cross-lingual SER will play a major role in future SER research to reduce the gap in the SER research between the well-resourced and under-resourced languages. Adaptation techniques such as MLLR and MAP may be useful in cross-lingual SER research.

While continuous and spectral features dominate and contribute to high recognition accuracy, there is room for researchers to examine the usefulness of qualitative and TEO-based features in SER research, since they not much been much considered over the past 12 years.

6 Conclusion

This paper has reviewed and analysed the existing research in SER, and found it to be an active field of research over the past 12 years, and doubtless one that will continue to be a focus of research in the future. Existing research was found to concentrate on several key issues including databases, classification, real time recognition, and cross-lingual SER that are still not fully resolved, and which are likely to be the focus of SER research in the future. Among the aspects of SER that could be taken into consideration in future works are the use of more appropriate databases, research on under resourced languages, real time recognition, and the inclusion of new types of emotion to fit the two dimensions valence and arousal. There are also long-term benefits to be obtained by integrating SER research more closely into the wider cross-disciplinary study of emotion including audio-visual, gesture and neurological aspect of human emotions, which can be very helpful for real-time recognition of emotion.

Funding The funding was provided by University of Malaya Research Grant (AFR (Frontier Science)) (Grant Number: RG284-14AFR), Postgraduate Research Grant (PPP) (Grant Number: PG220-2014B).

References

Abdelwahab, M., & Busso, C. (2017). Incremental adaptation using active learning for acoustic emotion Recognition. In *International conference on acoustics, speech and signal processing*.

Alam, M. J., Attabi, Y., Dumouchel, P., Kenny, P., & O'Shaughnessy, D. D. (2013). Amplitude modulation features for emotion recognition from speech. In *INTERSPEECH* (pp. 2420–2424).

Albornoz, E. M., Crolla, M. B., & Milone, D. H. (2008). Recognition of emotions in speech. In *Proceedings of XXXIV CLEI, Santa Fe Argentina*, pp. 1120–1129.

Albornoz, E. M., Milone, D. H., & Rufiner, H. L. (2011). Spoken emotion recognition using hierarchical classifiers. *Computer Speech & Language, 25*(3), 556–570.

Altun, H., & Polat, G. (2009). Boosting selection of speech related features to improve performance of multi-class SVMs in emotion detection. *Expert Systems with Applications, 36*(4), 8197–8203.

Álvarez, A., Cearreta, I., López, J. M., Arruti, A., Lazkano, E., Sierra, B., & Garay, N. (2007). A comparison using different speech parameters in the automatic emotion recognition using Feature Subset Selection based on Evolutionary Algorithms. In *International conference on text, speech and dialogue* (pp. 423–430). Berlin: Springer.

Anagnostopoulos, C. N., Iliou, T., & Giannoukos, I. (2012). Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review, 43*(2), 155–177.

Ananthakrishnan, S., Vembu, A. N., & Prasad, R. (2011). Model-based parametric features for emotion recognition from speech. In *2011 IEEE workshop on automatic speech recognition and understanding (ASRU)*, (pp. 529–534). Piscataway: IEEE.

Arias, J. P., Busso, C., & Yoma, N. B. (2013). Energy and F0 contour modeling with functional data analysis for emotional speech detection. In *INTERSPEECH* (pp. 2871–2875).

Arias, J. P., Busso, C., & Yoma, N. B. (2014). Shape-based modeling of the fundamental frequency contour for emotion detection in speech. *Computer Speech & Language, 28*(1), 278–294.

Atassi, H., & Esposito, A. (2008). A speaker independent approach to the classification of emotional vocal expressions. In *20th IEEE international conference on tools with artificial intelligence, 2008. ICTAI'08*. (Vol. 2, pp. 147–152). Piscataway: IEEE.

Atassi, H., Smekal, Z., & Esposito, A. (2012). Emotion recognition from spontaneous Slavic speech. In *2012 IEEE 3rd international conference on cognitive infocommunications* (*CogInfoCom*) (pp. 389–394). Piscataway: IEEE.

Athanaselis, T., Bakamidis, S., Dologlou, I., Cowie, R., Douglas-Cowie, E., & Cox, C. (2005). ASR for emotional speech: Clarifying the issues and enhancing performance. *Neural Networks, 18*(4), 437–444.

Attabi, Y., & Dumouchel, P. (2012). Emotion recognition from speech: WOC-NN and class-interaction. In *2012 11th international conference on information science, signal processing and their applications* (*ISSPA*) (pp. 126–131). Piscataway: IEEE.

Attabi, Y., & Dumouchel, P. (2013). Anchor models for emotion recognition from speech. *IEEE Transactions on Affective Computing, 4*(3), 280–290.

Bahreini, K., Nadolski, R., & Westera, W. (2016). Towards real-time speech emotion recognition for affective e-learning. *Education and Information Technologies, 21*(5), 1367–1386.

Balti, H., & Elmaghraby, A. S. (2013). Speech emotion detection using time dependent self organizing maps. In *2013 IEEE international symposium on signal processing and information technology* (*ISSPIT*) (pp. 000470–000478). Piscataway: IEEE.

Barra Chicote, R., Fernández Martínez, F., Lutfi, L., Binti, S., Lucas Cuesta, J. M., Macías Guarasa, J., … Pardo Muñoz, J. M. (2009). *Acoustic emotion recognition using dynamic Bayesian networks and multi-space distributions*. ISCA.

Batliner, A., Schuller, B., Seppi, D., Steidl, S., Devillers, L., Vidrascu, L., & Amir, N. (2011). The automatic recognition of emotions in speech. In *Emotion-oriented systems* (pp. 71–99). Berlin Heidelberg: Springer.

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Laskowski, K., Vogt, T., … Aharonson, V. (2006). Combining efforts for improving automatic classification of emotional user states. Proc. IS-LTC 240–245.

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Devillers, L., … Aharonson, V. (2007). The impact of F0 extraction errors on the classification of prominence and emotion. Proc. ICPhS 2201–2204.

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., & Rose, R. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication, 49*(10), 763–786.

Bertero, D., & Fung, P. (2017). A first look into a Convolutional Neural Network for speech emotion detection. In *2017 IEEE international conference on acoustics, speech and signal processing* (*ICASSP*) (pp. 5115–5119). Piscataway: IEEE.

Bhaykar, M., Yadav, J., & Rao, K. S. (2013). Speaker dependent, speaker independent and cross language emotion recognition from speech using GMM and HMM. *2013 national conference on communications (NCC)* (pp. 1–5). Piscataway: IEEE.

Bitouk, D., Nenkova, A., & Verma, R. (2009). Improving emotion recognition using class-level spectral features. In *INTERSPEECH* (pp. 2023–2026).

Böck, R., Hübner, D., & Wendemuth, A. (2010). Determining optimal signal features and parameters for hmm-based emotion classification. In *15th IEEE mediterranean electrotechnical conference MELECON 2010–2010* (pp. 1586–1590). Piscataway: IEEE.

Bojanić, M., Crnojević, V., & Delić, V. (2012). Application of neural networks in emotional speech recognition. In *2012 11th* *symposium on neural network applications in electrical engineering (NEUREL)* (pp. 223–226). Piscataway: IEEE.

Bozkurt, E., Erzin, E., Erdem, C. E., & Erdem, A. T. (2010). Use of line spectral frequencies for emotion recognition from speech. In *2010 20th international conference on pattern recognition* (*ICPR*) (pp. 3708–3711). Piscataway: IEEE.

Bozkurt, E., Erzin, E., Erdem, C. E., & Erdem, A. T. (2011). Formant position based weighted spectral features for emotion recognition. *Speech Communication, 53*(9), 1186–1197.

Bozkurt, E., Erzin, E., Eroğlu Erdem, Ç, & Erdem, T. (2009). Improving automatic emotion recognition from speech signals. In *10th annual conference of the international speech communication association 2009 (INTERSPEECH 2009)*. International Speech Communications Association.

Brester, C., Semenkin, E., & Sidorov, M. (2016). Multi-objective heuristic feature selection for speech-based multilingual emotion recognition. *Journal of Artificial Intelligence and Soft Computing Research, 6*(4), 243–253.

Brooks, C. A., Thompson, C., & Kovanović, V. (2016). Introduction to data mining for educational researchers. In *Proceedings of the 6th international conference on learning analytics & knowledge* (pp. 505–506). ACM.

Busso, C., Lee, S., & Narayanan, S. (2009). Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE Transactions on Audio, Speech, and Language Processing, 17*(4), 582–596.

Busso, C., Mariooryad, S., Metallinou, A., & Narayanan, S. (2013). Iterative feature normalization scheme for automatic emotion detection from speech. *IEEE Transactions on Affective Computing, 4*(4), 386–397.

Busso, C., Metallinou, A., & Narayanan, S. S. (2011). Iterative feature normalization for emotional speech detection. In *2011 IEEE international conference on acoustics, speech and signal processing* (*ICASSP*) (pp. 5692–5695). Piscataway: IEEE.

Calvo, R. A., & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing, 1*(1), 18–37.

Casale, S., Russo, A., Scebba, G., & Serrano, S. (2008). Speech emotion classification using machine learning algorithms. In *2008 IEEE international conference on semantic computing* (pp. 158–165). Piscataway: IEEE.

Casale, S., Russo, A., & Serrano, S. (2010). Analysis of robustness of attributes selection applied to speech emotion recognition. In *2010 18th European signal processing conference* (pp. 1174–1178). Piscataway: IEEE.

Chakraborty, R., Pandharipande, M., & Kopparapu, S. K. (2016). Knowledge-based framework for intelligent emotion recognition in spontaneous speech. *Procedia Computer Science, 96*, 587–596.

Chandaka, S., Chatterjee, A., & Munshi, S. (2009). Support vector machines employing cross-correlation for emotional speech recognition. *Measurement, 42*(4), 611–618.

Chandrakala, S., & Sekhar, C. C. (2009). Combination of generative models and SVM based classifier for speech emotion recognition. In *International joint conference on neural networks, 2009. IJCNN 2009* (pp. 497–502). Piscataway: IEEE.

Chavhan, Y., Dhore, M. L., & Yesaware, P. (2010). Speech emotion recognition using support vector machine. *International Journal of Computer Applications, 1*(20), 6–9.

Chavhan, Y. D., Yelure, B. S., & Tayade, K. N. (2015). Speech emotion recognition using RBF kernel of LIBSVM. In *2015 2nd international conference on electronics and communication systems* (*ICECS*) (pp. 1132–1135). Piscataway: IEEE.

Chen, L., Mao, X., Wei, P., Xue, Y., & Ishizuka, M. (2012). Mandarin emotion recognition combining acoustic and emotional point information. *Applied Intelligence, 37*(4), 602–612.

Chenchah, F., & Lachiri, Z. (2014). Speech emotion recognition in acted and spontaneous context. *Procedia Computer Science, 39*, 139–145.

Cheng, X., & Duan, Q. (2012). Speech emotion recognition using gaussian mixture model. In *The 2nd international conference on computer application and system modeling*.

Chiou, B. C., & Chen, C. P. (2013). Feature space dimension reduction in speech emotion recognition using support vector machine. In *Signal and information processing association annual summit and conference (APSIPA), 2013 Asia-Pacific* (pp. 1–6). Piscataway: IEEE.

Christina, I. J., & Milton, A. (2012). Analysis of all pole model to recognize emotions from speech signal. In *2012 international conference on computing, electronics and electrical technologies* (*ICCEET*) (pp. 723–728). Piscataway: IEEE.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schroder, M. (2000) FEELTRACE: an instrument for recording perceived emotion in real time. In *Proceedings of ISCA speech and emotion workshop*, pp 19–24.

Cummins, N., Amiriparian, S., Hagerer, G., Batliner, A., Steidl, S., & Schuller, B. (2017). An Image-based deep spectrum feature representation for the recognition of emotional speech. In *Proceedings of the 25th ACM international conference on multimedia, MM*. Piscataway: IEEE.

D'Mello, S., & Kory, J. (2012). Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies. In *Proceedings of the 14th ACM international conference on multimodal interaction* (pp. 31–38). ACM.

Dai, K., Fell, H. J., & MacAuslan, J. (2008). Recognizing emotion in speech using neural networks. *Telehealth and Assistive Technologies, 31*, 38–43.

Delic, V., Bojanic, M., Gnjatovic, M., Secujski, M., & Jovicic, S. T. (2012). Discrimination capability of prosodic and spectral features for emotional speech recognition. *Elektronika ir Elektrotechnika, 18*(9), 51–54.

Deng, J., Han, W., & Schuller, B. (2012). Confidence measures for speech emotion recognition: A start. In *Proceedings of speech communication; 10. ITG symposium* (pp. 1–4). VDE.

Deng, J., Xu, X., Zhang, Z., Frühholz, S., Grandjean, D., & Schuller, B. (2017). Fisher kernels on phase-based features for speech emotion recognition. In *Dialogues with social robots* (pp. 195–203). Springer: Singapore.

Deng, J., Zhang, Z., Eyben, F., & Schuller, B. (2014). Autoencoder-based unsupervised domain adaptation for speech emotion recognition. *IEEE Signal Processing Letters, 21*(9), 1068–1072.

Deng, J., Zhang, Z., Marchi, E., & Schuller, B. (2013). Sparse autoencoder-based feature transfer learning for speech emotion recognition. In *2013 humaine association conference on affective computing and intelligent interaction (ACII)* (pp. 511–516). Piscataway: IEEE.

Devillers, L., & Vidrascu, L. (2006). Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs. In *9th international conference on spoken language processing*.

Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. (2003). Emotional speech: towards a new generation of databases. *Speech Communication, 40*, 33–60.

Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., McRorie, M., Martin, J. C., Devillers, L., Abrilan, S., Batliner, A., Amir, N., & Karpouzis, K. (2007) The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data. In *Proceedings of international conference affective computing and intelligent interaction*, pp 488–500.

Ekman, P. (1957). A methodological discussion of non-verbal behavior. *Journal of Psychology, 43*, 141–149.

Ekman, P. (1972). Universals and cultural differences in facial expression of emotion. In J. Cole (Ed.), *Nebraska symposium on motivation* (pp. 207–283). Lincoln, NE: University of Nebraska Press.

Ekman, P. (1999). Basic emotions. In T. Dalgleish & M. Power (Eds.), *Handbook of cognition and emotion*. Chichester: Wiley.

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition, 44*(3), 572–587.

Elbarougy, R., & Akagi, M. (2012). Speech emotion recognition system based on a dimensional approach using a three-layered model. In *Signal & information processing association annual summit and conference (APSIPA ASC), 2012 Asia-Pacific* (pp. 1–9). Piscataway: IEEE.

Elbarougy, R., & Akagi, M. (2013). Cross-lingual speech emotion recognition system based on a three-layer model for human perception. In *Signal and information processing association annual summit and conference (APSIPA), 2013 Asia-Pacific* (pp. 1–10). Piscataway: IEEE.

Erdem, C. E., Bozkurt, E., Erzin, E., & Erdem, A. T. (2010). RANSAC-based training data selection for emotion recognition from spontaneous speech. In *Proceedings of the 3rd international workshop on affective interaction in natural environments* (pp. 9–14). ACM.

Esmaileyan, Z., & Marvi, H. (2014). Recognition of emotion in speech using variogram based features. *Malaysian Journal of Computer Science, 27*(3), 156–170.

Espinosa, H. P., García, C. A. R., & Pineda, L. V. (2010). Features selection for primitives estimation on emotional speech. In *2010 IEEE international conference on acoustics speech and signal processing* (*ICASSP*) (pp. 5138–5141). Piscataway: IEEE.

Fayek, H. M., Lech, M., & Cavedon, L. (2016). On the correlation and transferability of features between automatic speech recognition and speech emotion recognition. In *INTERSPEECH* (pp. 3618–3622).

Feraru, M., & Zbancioc, M. (2013). Speech emotion recognition for SROL database using weighted KNN algorithm. In *2013 international conference on electronics, computers and artificial intelligence* (*ECAI*) (pp. 1–4). Piscataway: IEEE.

Fernandez, R., & Picard, R. (2011). Recognizing affect from speech prosody using hierarchical graphical models. *Speech Communication, 53*(9), 1088–1103.

Firoz Shah, A., Vimal, K. V. R., Raji, S. A., Jayakumar, A., & Babu, A. P. (2009) Speaker independent automatic emotion recognition from speech: a comparison of MFCCs and discrete wavelet transforms. In *Proceedings of international conference on advances in recent technologies in communication and computing*, pp 528–531.

Fu, L., Mao, X., & Chen, L. (2008a). Relative speech emotion recognition based artificial neural network. In *Pacific-Asia workshop on computational intelligence and industrial application, 2008. PACIIA'08*. (Vol. 2, pp. 140–144). Piscataway: IEEE.

Fu, L., Mao, X., & Chen, L. (2008b). Speaker independent emotion recognition based on SVM/HMMs fusion system. In *International conference on audio, language and image processing, 2008. ICALIP 2008* (pp. 61–65). Piscataway: IEEE.

Gamage, K. W., Sethu, V., & Ambikairajah, E. (2017). Salience based lexical features for emotion recognition. In *2017 IEEE international conference on acoustics, speech and signal processing* (*ICASSP*) (pp. 5830–5834). Piscataway: IEEE.

Garg, V., Kumar, H., & Sinha, R. (2013). Speech based emotion recognition based on hierarchical decision tree with SVM, BLG and SVR classifiers. In *2013 national conference on communications (NCC)* (pp. 1–5). Piscataway: IEEE.

Gaurav, M. (2008). Performance analysis of spectral and prosodic features and their fusion for emotion recognition in speech. In *Spoken language technology workshop, 2008. SLT 2008* (pp. 313–316). Piscataway: IEEE.

Georgogiannis, A., & Digalakis, V. (2012). Speech emotion recognition using non-linear teager energy based features in noisy environments. In *2012 proceedings of the 20th European signal processing conference* (*EUSIPCO*) (pp. 2045–2049). Piscataway: IEEE.

Gharavian, D., Sheikhan, M., & Ashoftedel, F. (2013). Emotion recognition improvement using normalized formant supplementary features by hybrid of DTW-MLP-GMM model. *Neural Computing and Applications, 22*(6), 1181–1191.

Gharavian, D., Sheikhan, M., & Janipour, M. (2010). Pitch in emotional speech and emotional speech recognition using pitch frequency. *Majlesi Journal of Electrical Engineering*, 4(1).

Gharavian, D., Sheikhan, M., Nazerieh, A., & Garoucy, S. (2012). Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network. *Neural Computing and Applications, 21*(8), 2115–2126.

Gharsellaoui, S., Selouani, S. A., & Dahmane, A. O. (2015). Automatic emotion recognition using auditory and prosodic indicative features. In *2015 IEEE 28th Canadian conference on electrical and computer engineering (CCECE)* (pp. 1265–1270). Piscataway: IEEE.

Giannoulis, P., & Potamianos, G. (2012). A hierarchical approach with feature selection for emotion recognition from speech. In *LREC* (pp. 1203–1206).

Glüge, S., Böck, R., & Wendemuth, A. (2011). Segmented-memory recurrent neural networks versus hidden markov models in emotion recognition from speech. In *IJCCI (NCTA)* (pp. 308–315).

Grimm, M., Kroschel, K., Mower, E., & Narayanan, S. (2007a). Primitives-based evaluation and estimation of emotions in speech. *Speech Communication, 49*(10), 787–800.

Grimm, M., Kroschel, K., & Narayanan, S. (2007b). Support vector regression for automatic recognition of spontaneous emotions in speech. In *IEEE international conference on acoustics, speech and signal processing, 2007. ICASSP 2007*. (Vol. 4, pp. IV–1085). Piscataway: IEEE.

Hamidi, M., & Mansoorizade, M. (2012). Emotion recognition from persian speech with neural network. *International Journal of Artificial Intelligence & Applications, 3*(5), 107.

Han, J., Zhang, Z., Ringeval, F., & Schuller, B. (2017). Prediction-based learning for continuous emotion recognition in speech. In *42nd IEEE international conference on acoustics, speech, and signal processing, ICASSP 2017*.

Han, K., Yu, D., & Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. In *15th annual conference of the international speech communication association*.

Harimi, A., Fakhr, H. S., & Bakhshi, A. (2016). Recognition of emotion using reconstructed phase space of speech. *Malaysian Journal of Computer Science, 29*(4), 262–271.

Hassan, A., & Damper, R. I. (2009). Emotion recognition from speech using extended feature selection and a simple classifier. In *10th annual conference of the international speech communication association*.

He, L., Lech, M., Maddage, N. C., & Allen, N. B. (2011). Study of empirical mode decomposition and spectral analysis for stress and emotion classification in natural speech. *Biomedical Signal Processing and Control, 6*(2), 139–146.

Henríquez, P., Alonso, J. B., Ferrer, M. A., Travieso, C. M., & Orozco-Arroyave, J. R. (2014). Nonlinear dynamics characterization of emotional speech. *Neurocomputing, 132*, 126–135.

Hu, H., Xu, M. X., & Wu, W. (2007). Fusion of global statistical and segmental spectral features for speech emotion recognition. In *INTERSPEECH* (pp. 2269–2272).

Hu, H., Xu, M. X., & Wu, W. (2007). GMM supervector based SVM with spectral features for speech emotion recognition. In *IEEE international conference on acoustics, speech and signal processing, 2007. ICASSP 2007*. (Vol. 4, pp. IV–413). Piscataway: IEEE.

Huang, R., & Ma, C. (2006). Toward a speaker-independent real-time affect detection system. In *18th international conference on pattern recognition, 2006. ICPR 2006*. (Vol. 1, pp. 1204–1207). Piscataway: IEEE.

Huang, Y., Wu, A., Zhang, G., & Li, Y. (2016). Speech emotion recognition based on deep belief networks and wavelet packet cepstral coefficients. *International Journal of Simulation: Systems, Science and Technology, 17*(28), 28–31.

Huang, Z., Dong, M., Mao, Q., & Zhan, Y. (2014). Speech emotion recognition using CNN. In *Proceedings of the 22nd ACM international conference on Multimedia* (pp. 801–804). ACM.

Hussain, L., Shafi, I., Saeed, S., Abbas, A., Awan, I. A., Nadeem, S. A., … Rahman, B. (2017). A radial base neural network approach for emotion recognition in human speech. *IJCSNS, 17*(8), 52.

Iliev, A. I., & Scordilis, M. S. (2008). Emotion recognition in speech using inter-sentence Glottal statistics. In *15th international conference on systems, signals and image processing, 2008. IWSSIP 2008*. (pp. 465–468). Piscataway: IEEE.

Iliev, A. I., Scordilis, M. S., Papa, J. P., & Falcão, A. X. (2010). Spoken emotion recognition through optimum-path forest classification using glottal features. *Computer Speech & Language, 24*(3), 445–460.

Iliou, T., & Anagnostopoulos, C. N. (2009). Comparison of different classifiers for emotion recognition. In *13th panhellenic conference on informatics, 2009. PCI'09*. (pp. 102–106). Piscataway: IEEE.

Iliou, T., & Anagnostopoulos, C. N. (2010a). SVM-MLP-PNN classifiers on speech emotion recognition field—A comparative study. In *2010 fifth international conference on digital telecommunications* (*ICDT*) (pp. 1–6). Piscataway: IEEE.

Iliou, T., & Anagnostopoulos, C. N. (2010b). Classification on speech emotion recognition-a comparative study. *Animation, 4*, 5.

Iriondo, I., Planet, S., Alías, F., Socoró, J. C., & Martínez, E. (2007). Validation of an expressive speech corpus by mapping automatic classification to subjective evaluation. *Computational and Ambient Intelligence*, 646–653.

Ivanov, A., & Riccardi, G. (2012). Kolmogorov-Smirnov test for feature selection in emotion recognition from speech. In *2012 IEEE international conference on acoustics, speech and signal processing* (*ICASSP*) (pp. 5125–5128). Piscataway: IEEE.

Javidi, M. M., & Roshan, E. F. (2013). Speech emotion recognition by using combinations of C5. 0, neural network (NN), and support vector machines (SVM) classification methods. *International Journal of Applied Mathematics and Computer Science, 6*, 191–200.

Jeon, J. H., Xia, R., & Liu, Y. (2011). Sentence level emotion recognition based on decisions from subsentence segments. In *2011 IEEE international conference on acoustics, speech and signal processing* (*ICASSP*) (pp. 4940–4943). Piscataway: IEEE.

Jiang, J., Wu, Z., Xu, M., Jia, J., & Cai, L. (2012). Comparison of adaptation methods for GMM-SVM based speech emotion recognition. In *2012 IEEE spoken language technology workshop (SLT)* (pp. 269–273). Piscataway: IEEE.

Jin, Q., Li, C., Chen, S., & Wu, H. (2015). Speech emotion recognition with acoustic and lexical features. In *2015 IEEE international conference on acoustics, speech and signal processing* (*ICASSP*) (pp. 4749–4753). Piscataway: IEEE.

Kamińska, D., & Pelikant, A. (2012). Recognition of human emotion from a speech signal based on Plutchik's model. *International Journal of Electronics and Telecommunications, 58*(2), 165–170.

Kandali, A. B., Routray, A., & Basu, T. K. (2008). Emotion recognition from Assamese speeches using MFCC features and GMM classifier. In *TENCON 2008–2008 IEEE region 10 conference* (pp. 1–5). Piscataway: IEEE.

Khan, M., Goskula, T., Nasiruddin, M., & Quazi, R. (2011). Comparison between k-nn and svm method for speech emotion recognition. *International Journal on Computer Science and Engineering, 3*(2), 607–611.

Khanna, P., & Kumar, M. S. (2011). Application of vector quantization in emotion recognition from human speech. In *International conference on information intelligence, systems, technology and management* (pp. 118–125). Berlin, Heidelberg: Springer.

Kim, E. H., Hyun, K. H., Kim, S. H., & Kwak, Y. K. (2009). Improved emotion recognition with a novel speaker-independent feature. *IEEE/ASME Transactions on Mechatronics, 14*(3), 317–325.

Kim, E. H., Hyun, K. H., & Kwak, Y. K. (2006). Improvement of emotion recognition from voice by separating of obstruents. In *The 15th IEEE international symposium on robot and human interactive communication, 2006. ROMAN 2006*. (pp. 564–568). Piscataway: IEEE.

Kim, J. B., Park, J. S., & Oh, Y. H. (2011). On-line speaker adaptation based emotion recognition using incremental emotional information. In *2011 IEEE international conference on acoustics, speech and signal processing* (*ICASSP*) (pp. 4948–4951). Piscataway: IEEE.

Kim, J. B., Park, J. S., & Oh, Y. H. (2012). Speaker-characterized emotion recognition using online and iterative speaker adaptation. *Cognitive Computation, 4*(4), 398–408.

Kim, S., Georgiou, P. G., Lee, S., & Narayanan, S. (2007). Real-time emotion detection system using speech: Multi-modal fusion of different timescale features. In *IEEE 9th workshop on multimedia signal processing, 2007. MMSP 2007* (pp. 48–51).

Kishore, K. K., & Satish, P. K. (2013). Emotion recognition in speech using MFCC and wavelet features. In *2013 IEEE 3rd international advance computing conference (IACC)* (pp. 842–847). Piscataway: IEEE.

Kitchenham, B. (2004). Procedures for performing systematic reviews. Keele, Keele University 33.

Koolagudi, S. G., & Krothapalli, R. S. (2011). Two stage emotion recognition based on speaking rate. *International Journal of Speech Technology, 14*(1), 35–48.

Koolagudi, S. G., & Krothapalli, S. R. (2012). Emotion recognition from speech using sub-syllabic and pitch synchronous spectral features. *International Journal of Speech Technology, 15*(4), 495–511.

Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech using source, system, and prosodic features. *International Journal of Speech Technology, 15*(2), 265–289.

Koolagudi, S. G., Reddy, R., & Rao, K. S. (2010). Emotion recognition from speech signal using epoch parameters. In *2010 international conference on signal processing and communications* (*SPCOM*) (pp. 1–5). Piscataway: IEEE.

Kostoulas, T., Ganchev, T., Lazaridis, A., & Fakotakis, N. (2010) Enhancing Emotion recognition from speech through feature selection. In P. Sojka, A. Horák, I. Kopecek & K. Pala (Eds.) *Text, speech and dialogue, lecture notes in artificial intelligence*, Vol. 6231, pp. 338–344.

Kostoulas, T., Ganchev, T., Mporas, I., & Fakotakis, N. (2007) Detection of negative emotional states in real-world scenario. In *Proceedings of 19th IEEE international conference on tools with artificial intelligence*, pp 502–509.

Kotti, M., Paterno, F., & Kotropoulos, C. (2010). Speaker-independent negative emotion recognition. In *2010 2nd international workshop on cognitive information processing* (*CIP*) (pp. 417–422). Piscataway: IEEE.

Le, D., Aldeneh, Z., & Provost, E. M. (2017). Discretized continuous speech emotion recognition with multi-task deep recurrent neural network. *Interspeech*, 2017.

Le, D., & Provost, E. M. (2013). Emotion recognition from spontaneous speech using hidden markov models with deep belief networks. In *2013 IEEE workshop on automatic speech recognition and understanding (ASRU)* (pp. 216–221). Piscataway: IEEE.

Lee, J., & Tashev, I. (2015). High-level feature representation using recurrent neural network for speech emotion recognition. In *INTERSPEECH* (pp. 1537–1540).

Lefter, I., Rothkrantz, L. J., Wiggers, P., & Van Leeuwen, D. A. (2010). Emotion recognition from speech by combining databases and fusion of classifiers. In *Text, speech and dialogue* (pp. 353–360). Berlin Heidelberg: Springer.

Li, L., Zhao, Y., Jiang, D., Zhang, Y., Wang, F., Gonzalez, I., … Sahli, H. (2013). Hybrid deep neural network–hidden markov model (DNN-HMM) based speech emotion recognition. In *2013 humaine association conference on affective computing and intelligent interaction (ACII)* (pp. 312–317). Piscataway: IEEE.

Li, Y., Chao, L., Liu, Y., Bao, W., & Tao, J. (2015) From simulated speech to natural speech, what are the robust features for emotion recognition? In *International conference on affective computing and intelligent interaction (ACII)* (pp. 368–373). Piscataway: IEEE

Lim, W., Jang, D., & Lee, T. (2016). Speech emotion recognition using convolutional and recurrent neural networks. In *Signal and information processing association annual summit and conference (APSIPA), 2016 Asia-Pacific* (pp. 1–4). Piscataway: IEEE.

Litman, D. J., & Forbes-Riley, K. (2006). Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. *Speech Communication, 48*(5), 559–590.

Liu, J., Chen, C., Bu, J., You, M., & Tao, J. (2007). Speech emotion recognition based on a fusion of all-class and pairwise-class feature selection. *Computational Science–ICCS 2007*, 168–175.

Luengo, I., Navas, E., & Hernáez, I. (2010). Feature analysis and evaluation for automatic emotion identification in speech. *IEEE Transactions on Multimedia, 12*(6), 490–501.

Lugger, M., Janoir, M. E., & Yang, B. (2009). Combining classifiers with diverse feature sets for robust speaker independent emotion recognition. In *Signal processing conference, 2009 17th European* (1225–1229). Piscataway: IEEE.

Lugger, M., & Yang, B. (2007). The relevance of voice quality features in speaker independent emotion recognition. In *IEEE international conference on acoustics, speech and signal processing, 2007. ICASSP 2007.* (Vol. 4, pp. IV–17). Piscataway: IEEE.

Lugger, M., & Yang, B. (2007). An incremental analysis of different feature groups in speaker independent emotion recognition. In *16th Int. congress of phonetic sciences*.

Mannepalli, K., Sastry, P. N., & Suman, M. (2016). A novel adaptive fractional deep belief networks for speaker emotion recognition. *Alexandria Engineering Journal*.

Mao, Q., Xue, W., Rao, Q., Zhang, F., & Zhan, Y. (2016). Domain adaptation for speech emotion recognition by sharing priors between related source and target classes. In *2016 IEEE international conference on acoustics, speech and signal processing* (*ICASSP*) (pp. 2608–2612). Piscataway: IEEE.

Mao, X., Chen, L., & Fu, L. (2009). Multi-level speech emotion recognition based on HMM and ANN. In *2009 WRI World congress on computer science and information engineering* (Vol. 7, pp. 225–229). Piscataway: IEEE.

Mao, X., Zhang, B., & Luo, Y. (2007). Speech emotion recognition based on a hybrid of HMM/ANN. In *Proceedings of the 7th conference on 7th WSEAS international conference on applied informatics and communications* (Vol. 7, pp. 367–370).

Mencattini, A., Martinelli, E., Ringeval, F., Schuller, B., & Di Natlae, C. (2017). Continuous estimation of emotions in speech by dynamic cooperative speaker models. In *IEEE transactions on affective computing*.

Milton, A., Roy, S. S., & Selvi, S. T. (2013). Svm scheme for speech emotion recognition using mfcc feature. *International Journal of Computer Applications*, 69(9).

Milton, A., & Selvi, S. T. (2014). Class-specific multiple classifiers scheme to recognize emotions from speech signals. *Computer Speech & Language, 28*(3), 727–742.

Mishra, H. K., & Sekhar, C. C. (2009). Variational Gaussian mixture models for speech emotion recognition. In *Seventh international conference on advances in pattern recognition, 2009. ICAPR'09*. (pp. 183–186). Piscataway: IEEE.

Morales-Perez, M., Echeverry-Correa, J., Orozco-Gutierrez, A., & Castellanos-Dominguez, G. (2008). Feature extraction of speech signals in emotion identification. In *Engineering in medicine and biology society, 2008. EMBS 2008. 30th annual international conference of the IEEE* (pp. 2590–2593). Piscataway: IEEE.

Morrison, D., Wang, R., & De Silva, L. C. (2007). Ensemble methods for spoken emotion recognition in call-centres. *Speech communication, 49*(2), 98–112.

Navas, E., Hernáez, I., & Luengo, I. (2006). An objective and subjective study of the role of semantics and prosodic features in building corpora for emotional TTS. *EEE Transactions on Audio, Speech and Language Processing 14*, 1117–1127.

Neiberg, D., & Elenius, K. (2008). Automatic recognition of anger in spontaneous speech. In *9th annual conference of the international speech communication association*.

Nicolaou, M. A., Gunes, H., & Pantic, M. (2011). Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing, 2*(2), 92–105.

Ntalampiras, S., & Fakotakis, N. (2012). Modeling the temporal evolution of acoustic parameters for speech emotion recognition. *IEEE Transactions on Affective Computing, 3*(1), 116–125.

Pan, Y., Shen, P., & Shen, L. (2012). Speech emotion recognition using support vector machine. *International Journal of Smart Home, 6*(2), 101–108.

Pao, T. L., Chien, C. S., Chen, Y. T., Yeh, J. H., Cheng, Y. M., & Liao, W. Y. (2007). Combination of multiple classifiers for improving emotion recognition in Mandarin speech. In *3rd international conference on intelligent information hiding and multimedia signal processing, 2007. IIHMSP 2007* (Vol. 1, pp. 35–38). Piscataway: IEEE.

Pao, T. L., Wang, C. H., & Li, Y. J. (2012). A study on the search of the most discriminative speech features in the speaker dependent speech emotion recognition. In *2012 fifth international symposium on parallel architectures, algorithms and programming* (*PAAP*) (pp. 157–162). Piscataway: IEEE.

Pathak, S., & Kulkarni, A. (2011). Recognizing emotions from speech. In *2011 3rd international conference on electronics computer technology* (*ICECT*) (Vol. 4, pp. 107–109). Piscataway: IEEE.

Philippou-Hübner, D., Vlasenko, B., Böck, R., & Wendemuth, A. (2012). The performance of the speaking rate parameter in emotion recognition from speech. In *2012 IEEE international conference on multimedia and expo* (*ICME*) (pp. 248–253). Piscataway: IEEE.

Picard, R. W., & Picard, R. (1997). *Affective computing (252)*. Cambridge: MIT press.

Pierre-Yves, O. (2003). The production and recognition of emotions in speech: features and algorithms. *International Journal of Human-Computer Studies, 59*(1), 157–183.

Planet, S., & Iriondo, I. (2012). Comparison between decision-level and feature-level fusion of acoustic and linguistic features for spontaneous emotion recognition. In *2012 7th Iberian conference on information systems and technologies (CISTI)* (pp. 1–6). Piscataway: IEEE.

Plutchik, R. (1991). *The emotions*. Lanham, MD: University Press of America.

Pohjalainen, J., Fabien Ringeval, F., Zhang, Z., & Schuller, B. (2016). Spectral and cepstral audio noise reduction techniques in speech emotion recognition. In *Proceedings of the 2016 ACM on multimedia conference* (pp. 670–674). ACM.

Polzehl, T., Schmitt, A., Metze, F., & Wagner, M. (2011). Anger recognition in speech using acoustic and linguistic cues. *Speech Communication, 53*(9), 1198–1209.

Přibil, J., & Přibilová, A. (2013). Evaluation of influence of spectral and prosodic features on GMM classification of Czech and Slovak emotional speech. *EURASIP Journal on Audio, Speech, and Music Processing, 2013*(1), 8.

Rao, K. S., Koolagudi, S. G., & Vempada, R. R. (2013). Emotion recognition from speech using global and local prosodic features. *International Journal of Speech Technology, 16*(2), 143–160.

Rao, K. S., Kumar, T. P., Anusha, K., Leela, B., Bhavana, I., & Gowtham, S. V. S. K. (2012). Emotion recognition from speech. *International Journal of Computer Science and Information Technologies, 3*(2), 3603–3607.

Rehmam, B., Halim, Z., Abbas, G., & Muhammad, T. (2015). Artificial neural network-based speech recognition using Dwt analysis applied on isolated words from oriental languages. *Malaysian Journal of Computer Science, 28*(3), 242–262.

Ringeval, F., & Chetouani, M. (2008). Exploiting a vowel based approach for acted emotion recognition. In *Verbal and nonverbal features of human-human and human-machine interaction*, pp. 243–254.

Rodríguez, P. H., Hernández, J. B. A., Ballester, M. A. F., González, C. M. T., & Orozco-Arroyave, J. R. (2013). Global selection of features for nonlinear dynamics characterization of emotional speech. *Cognitive Computation, 5*(4), 517–525.

Rong, J., Li, G., & Chen, Y. P. P. (2009). Acoustic feature selection for automatic emotion recognition from speech. *Information Processing & Management, 45*(3), 315–328.

Sagha, H., Deng, J., Gavryukova, M., Han, J., & Schuller, B. (2016). Cross lingual speech emotion recognition using canonical correlation analysis on principal component subspace. In *2016 IEEE international conference on acoustics, speech and signal processing* (*ICASSP*) (pp. 5800–5804). Piscataway: IEEE.

Sánchez-Gutiérrez, M. E., Albornoz, E. M., Martinez-Licona, F., Rufiner, H. L., & Goddard, J. (2014). Deep learning for emotional speech recognition. In *Mexican conference on pattern recognition* (pp. 311–320). Cham: Springer International Publishing.

Scherer, S., Schwenker, F., & Palm, G. (2008). Emotion recognition from speech using multi-classifier systems and rbf-ensembles. In *Speech, audio, image and biomedical signal processing using neural networks*, pp. 49–70.

Scherer, S., Schwenker, F., & Palm, G. (2008). Emotion recognition from speech using multi-classifier systems and rbf-ensembles. In *Speech, audio, image and biomedical signal processing using neural networks* (pp. 49–70). Berlin Heidelberg: Springer.

Scherer, S., Schwenker, F., & Palm, G. (2008). Emotion recognition from speech using multi-classifier systems and rbf-ensembles. In *Speech, audio, image and biomedical signal processing using neural networks*, 49–70.

Scherer, S., Schwenker, F., & Palm, G. (2009). Classifier fusion for emotion recognition from speech. In *Advanced intelligent environments* (pp. 95–117). Springer US.

Schmitt, M., Ringeval, F., & Schuller, B. W. (2016). At the border of acoustics and linguistics: bag-of-audio-words for the recognition of emotions in speech. In *INTERSPEECH* (pp. 495–499).

Schuller, B., Seppi, D., Batliner, A., Maier, A., & Steidl, S. (2007). Towards more reality in the recognition of emotional speech. In *2007 IEEE international conference on acoustics, speech and signal processing-ICASSP'07* (Vol. 4, pp. IV–941). Piscataway: IEEE.

Schuller, B., Vlasenko, B., Eyben, F., Wollmer, M., Stuhlsatz, A., Wendemuth, A., & Rigoll, G. (2010). Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing, 1*(2), 119–131.

Schuller, B., Vlasenko, B., Minguez, R., Rigoll, G., & Wendemuth, A. (2007). Comparing one and two-stage acoustic modeling in the recognition of emotion in speech. In *IEEE workshop on automatic speech recognition & understanding, 2007. ASRU* (pp. 596–600). Piscataway: IEEE.

Schuller, B. W. (2008). Speaker, noise, and acoustic space adaptation for emotion recognition in the automotive environment. In *2008 ITG conference on voice communication (SprachKommunikation)* (pp. 1–4). VDE.

Schwenker, F., Scherer, S., Magdi, Y. M., & Palm, G. (2009). The GMM-SVM supervector approach for the recognition of the emotional status from speech. In *International conference on artificial neural networks* (pp. 894–903). Berlin, Heidelberg: Springer.

Sedaaghi, M. H., Kotropoulos, C., & Ververidis, D. (2007). Using adaptive genetic algorithms to improve speech emotion recognition. In *IEEE 9th workshop on multimedia signal processing, 2007. MMSP 2007*. (pp. 461–464). Piscataway: IEEE.

Seehapoch, T., & Wongthanavasu, S. (2013). Speech emotion recognition using support vector machines. In *2013 5th international conference on knowledge and smart technology* (*KST*) (pp. 86–91). Piscataway: IEEE.

Ser, W., Cen, L., & Yu, Z. L. (2008). A hybrid PNN-GMM classification scheme for speech emotion recognition. In *19th international conference on pattern recognition, 2008. ICPR 2008* (pp. 1–4). Piscataway: IEEE.

Sethu, V., Ambikairajah, E., & Epps, J. (2007). Speaker normalisation for speech-based emotion detection. In *2007 15th international conference on digital signal processing* (pp. 611–614). Piscataway: IEEE.

Sethu, V., Ambikairajah, E., & Epps, J. (2008a). Phonetic and speaker variations in automatic emotion classification. In *9th annual conference of the international speech communication association*.

Sethu, V., Ambikairajah, E., & Epps, J. (2008b). Empirical mode decomposition based weighted frequency feature for speech-based emotion classification. In *IEEE international conference on acoustics, speech and signal processing, 2008. ICASSP 2008*. (pp. 5017–5020). Piscataway: IEEE.

Sethu, V., Ambikairajah, E., & Epps, J. (2009). Speaker dependency of spectral features and speech production cues for automatic emotion classification. In *IEEE international conference on acoustics, speech and signal processing, 2009. ICASSP 2009*. (pp. 4693–4696). Piscataway: IEEE.

Sethu, V., Ambikairajah, E., & Epps, J. (2013). On the use of speech parameter contours for emotion recognition. *EURASIP Journal on Audio, Speech, and Music Processing, 2013*(1), 19.

Shah, F. (2009). Automatic emotion recognition from speech using artificial neural networks with gender-dependent databases. In *International conference on advances in computing, control, & telecommunication technologies, 2009. ACT'09*. (pp. 162–164). Piscataway: IEEE.

Shah, M., Miao, L., Chakrabarti, C., & Spanias, A. (2013). A speech emotion recognition framework based on latent Dirichlet allocation: Algorithm and FPGA implementation. In *2013 IEEE* *international conference on acoustics, speech and signal processing* (*ICASSP*) (pp. 2553–2557). Piscataway: IEEE.

Shami, M., & Verhelst, W. (2007). An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. *Speech Communication, 49*(3), 201–212.

Shaukat, A., & Chen, K. (2011). Emotional state recognition from speech via soft-competition on different acoustic representations. In *The 2011 international joint conference on neural networks (IJCNN)* (pp. 1910–1917). Piscataway: IEEE.

Shaw, A., Vardhan, R. K., & Saxena, S. (2016). Emotion recognition and classification in speech using Artificial neural networks. *International Journal of Computer Applications*, 145(8).

Sheikhan, M., Bejani, M., & Gharavian, D. (2013). Modular neural-SVM scheme for speech emotion recognition using ANOVA feature selection method. *Neural Computing and Applications, 23*(1), 215–227.

Sheikhan, M., Gharavian, D., & Ashoftedel, F. (2012). Using DTW neural–based MFCC warping to improve emotional speech recognition. *Neural Computing and Applications, 21*(7), 1765–1773.

Shen, P., Changjun, Z., & Chen, X. (2011). Automatic speech emotion recognition using support vector machine. In *2011 international conference on electronic and mechanical engineering and information technology* (*EMEIT*) (Vol. 2, pp. 621–625). Piscataway: IEEE.

Sidorov, M., Ultes, S., & Schmitt, A. (2014). Emotions are a personal thing: Towards speaker-adaptive emotion recognition. In *2014 IEEE international conference on acoustics, speech and signal processing* (*ICASSP*) (pp. 4803–4807). Piscataway: IEEE.

Soltani, K., & Ainon, R. N. (2007). Speech emotion detection based on neural networks. In *9th international symposium on signal processing and its applications, 2007. ISSPA 2007.* (pp. 1–3). Piscataway: IEEE.

Song, P., Ou, S., Zheng, W., Jin, Y., & Zhao, L. (2016). Speech emotion recognition using transfer non-negative matrix factorization. In *2016 IEEE international conference on acoustics, speech and signal processing* (*ICASSP*) (pp. 5180–5184). Piscataway: IEEE.

Song, P., Zheng, W., Ou, S., Zhang, X., Jin, Y., Liu, J., & Yu, Y. (2016). Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization. *Speech Communication, 83*, 34–41.

Steidl, S., Batliner, A., Nöth, E., & Hornegger, J. (2008). Quantification of segmentation and F0 errors and their effect on emotion recognition. In *Text, speech and dialogue* (pp. 525–534). Berlin/Heidelberg: Springer.

Sun, Y., & Wen, G. (2015). Emotion recognition using semi-supervised feature selection with speaker normalization. *International Journal of Speech Technology, 18*(3), 317–331.

Sun, Y., Wen, G., & Wang, J. (2015). Weighted spectral features based on local Hu moments for speech emotion recognition. *Biomedical Signal Processing and Control, 18*, 80–90.

Sun, Y., Zhou, Y., Zhao, Q., & Yan, Y. (2009). Acoustic feature optimization for emotion affected speech recognition. In *International conference on information engineering and computer science, 2009. ICIECS 2009.* (pp. 1–4). Piscataway: IEEE.

Swain, M., Sahoo, S., Routray, A., Kabisatpathy, P., & Kundu, J. N. (2015). Study of feature combination using HMM and SVM for multilingual Odiya speech emotion recognition. *International Journal of Speech Technology, 18*(3), 387–393.

Sztahó, D., Imre, V., & Vicsi, K. (2011). Automatic classification of emotions in spontaneous speech. Analysis of verbal and nonverbal communication and enactment. The Processing Issues, pp. 229–239.

Tabatabaei, T. S., Krishnan, S., & Guergachi, A. (2007). Emotion recognition using novel speech signal features. In *IEEE international symposium on circuits and systems, 2007. ISCAS 2007* (pp. 345–348). Piscataway: IEEE.

Tahon, M., & Devillers, L. (2015). Towards a small set of robust acoustic features for emotion recognition: IEEE/ACM transactions on challenges audio, speech, and language processing, *24*(1), 16–28.

Tamulevicius, G., & Liogiene, T. (2015). Low-order multi-level features for speech emotions recognition. *Baltic Journal of Modern Computing, 3*(4), 234–247.

Tarasov, A., & Delany, S. J. (2011). Benchmarking classification models for emotion recognition in natural speech: A multi-corporal study. In *2011 IEEE international conference on automatic face & gesture recognition and workshops* (*FG 2011*) (pp. 841–846). Piscataway: IEEE.

Ten Bosch, L. (2003). Emotions, speech and the ASR framework. *Speech Communication, 40*(1), 213–225.

Thapliyal, N., & Amoli, G. (2012). Speech based emotion recognition with gaussian mixture model. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 1*(5), 65.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2016). Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE international conference on acoustics, speech and signal processing* (*ICASSP*), (pp. 5200–5204). Piscataway: IEEE.

Truong, K., & Van Leeuwen, D. (2007). An 'open-set'detection evaluation methodology for automatic emotion recognition in speech. In *Workshop on paralinguistic speech-between models and data* (pp. 5–10).

Tseng, M., Hu, Y., Han, W. W., & Bergen, B. (2005). "Searching for happiness" or" Full of Joy"? Source domain activation matters. In *annual meeting of the Berkeley linguistics society* (Vol. 31, No. 1, pp. 359–370).

Utane, A. S., & Nalbalwar, S. L. (2013). Emotion recognition through speech using gaussian mixture model and support vector machine. *Emotion, 2*, 8.

Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication, 48*(9), 1162–1181.

Vlasenko, B., Philippou-Hübner, D., Prylipko, D., Böck, R., Siegert, I., & Wendemuth, A. (2011a). Vowels formants analysis allows straightforward detection of high arousal emotions. In *2011 IEEE international conference on multimedia and expo* (*ICME*) (pp. 1–6). Piscataway: IEEE.

Vlasenko, B., Prylipko, D., Philippou-Hübner, D., & Wendemuth, A. (2011b). Vowels formants analysis allows straightforward detection of high arousal acted and spontaneous emotions. In *12th annual conference of the international speech communication association*.

Vlasenko, B., Schuller, B., Wendemut, A., & Rigoll, G. (2007) Frame vs Turn-level: emotion recognition from speech considering static and dynamic processing. In *Proceedings 2nd international conference on affective computing and intelligent interaction*, pp 139–147.

Vogt, T., & André, E. (2006). Improving automatic emotion recognition from speech via gender differentiation. In *Proceeding language resources and evaluation conference (LREC 2006)*, Genoa.

Vogt, T., & André, E. (2009). Exploring the benefits of discretization of acoustic features for speech emotion recognition. In *10th annual conference of the international speech communication association*.

Vogt, T., & André, E. (2011). An evaluation of emotion units and feature types for real-time speech emotion recognition. *KI-Künstliche Intelligenz, 25*(3), 213–223.

Vondra, M., & Vích, R. (2009). Evaluation of speech emotion classification based on GMM and data fusion. In *Cross-modal analysis of speech, gestures, gaze and facial expressions*, pp. 98–105.

Wagner, J., Vogt, T., & André, E. (2007). A systematic comparison of different HMM designs for emotion recognition from acted and spontaneous speech. In *international conference on affective computing and intelligent interaction* (pp. 114–125). Springer, Berlin, Heidelberg.

Wang, F., Verhelst, W., & Sahli, H. (2011). Relevance vector machine based speech emotion recognition. In *Affective computing and intelligent interaction*, pp. 111–120.

Weninger, F., Ringeval, F., Marchi, E., & Schuller, B. W. (2016). Discriminatively trained recurrent neural networks for continuous dimensional emotion recognition from audio. In *IJCAI* (pp. 2196–2202).

Wenjing, H., Haifeng, L., & Chunyu, G. (2009). A hybrid speech emotion perception method of VQ-based feature processing and ANN recognition. In *WRI global congress on intelligent systems, 2009. GCIS'09*. (Vol. 2, pp. 145–149). Piscataway: IEEE.

Womack, B. D., & Hansen, J. H. (1999). N-channel hidden Markov models for combined stressed speech classification and recognition. *IEEE Transactions on Speech and Audio Processing, 7*(6), 668–677.

Wu, C. H., & Liang, W. B. (2011). Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Transaction Affective Computing, 2*, 10–21.

Wu, S., Falk, T. H., & Chan, W. Y. (2009). Automatic recognition of speech emotion using long-term spectro-temporal features. In *2009 16th international conference on digital signal processing* (pp. 1–6). Piscataway: IEEE.

Wu, S., Falk, T. H., & Chan, W. Y. (2011). Automatic speech emotion recognition using modulation spectral features. *Speech Communication, 53*(5), 768–785.

Wu, T., Yang, Y., Wu, Z., & Li, D. (2006). MASC: A speech corpus in mandarin for emotion analysis and affective speaker recognition. In *Speaker and language recognition workshop, 2006. IEEE Odyssey 2006* (pp. 1–5). Piscataway: IEEE.

Xiao, Z., Dellandréa, E., Chen, L., & Dou, W. (2009). Recognition of emotions in speech by a hierarchical approach. In *3rd international conference on affective computing and intelligent interaction and workshops, 2009. ACII 2009*. (pp. 1–8). Piscataway: IEEE.

Xiao, Z., Dellandrea, E., Dou, W., & Chen, L. (2006). Two-stage classification of emotional speech. In *international conference on digital telecommunications, 2006. ICDT'06*. (pp. 32–32). Piscataway: IEEE.

Xiao, Z., Dellandrea, E., Dou, W., & Chen, L. (2007, December). Automatic hierarchical classification of emotional speech. In *9th IEEE international symposium on multimedia workshops, 2007. ISMW'07*. (pp. 291–296). Piscataway: IEEE.

Xiao, Z., Dellandrea, E., Dou, W., & Chen, L. (2007). Hierarchical classification of emotional speech. *IEEE Transactions on Multimedia*, 37.

Yang, B., & Lugger, M. (2010). Emotion recognition from speech signals using new harmony features. *Signal Processing, 90*(5), 1415–1423.

Yang, N., Muraleedharan, R., Kohl, J., Demirkol, I., Heinzelman, W., & Sturge-Apple, M. (2012). Speech-based emotion classification using multiclass SVM with hybrid kernel and thresholding fusion. In *Spoken Language Technology Workshop (SLT), 2012 IEEE* (pp. 455–460). Piscataway: IEEE.

Ye, C., Liu, J., Chen, C., Song, M., & Bu, J. (2008). Speech emotion classification on a Riemannian manifold. In *Advances in multimedia information processing-PCM 2008*, pp. 61–69.

Yeh, J. H., Pao, T. L., Lin, C. Y., Tsai, Y. W., & Chen, Y. T. (2011). Segment-based emotion recognition from continuous Mandarin Chinese speech. *Computers in Human Behavior, 27*(5), 1545–1552.

You, M., Chen, C., Bu, J., Liu, J., & Tao, J. (2006). A hierarchical framework for speech emotion recognition. In *2006 IEEE international symposium on industrial electronics* (Vol. 1, pp. 515–519). Piscataway: IEEE.

You, M., Chen, C., Bu, J., Liu, J., & Tao, J. (2006). Emotional speech analysis on nonlinear manifold. In *18th international conference on pattern recognition, 2006. ICPR 2006*. (Vol. 3, pp. 91–94). Piscataway: IEEE.

Yun, S., & Yoo, C. D. (2012). Loss-scaled large-margin Gaussian mixture models for speech emotion classification. *IEEE Transactions on Audio, Speech, and Language Processing, 20*(2), 585–598.

Yüncü, E., Hacihabiboglu, H., & Bozsahin, C. (2014). Automatic speech emotion recognition using auditory models with binary decision tree and svm. In *2014 22nd international conference on pattern recognition* (*ICPR*) (pp. 773–778). Piscataway: IEEE.

Zbancioc, M., & Feraru, S. M. (2012). Emotion recognition of the SROL Romanian database using fuzzy KNN algorithm. In *10th international symposium on electronics and telecommunications* (*ISETC*), *2012* (pp. 347–350). Piscataway: IEEE.

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A survey of affect recognition methods: audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*, 39–58.

Zha, C., Yang, P., Zhang, X., & Zhao, L. (2016). Spontaneous speech emotion recognition via multiple kernel learning. In *2016 eighth international conference on measuring technology and mechatronics automation* (*ICMTMA*) (pp. 621–623). Piscataway: IEEE.

Zhang, S., Lei, B., Chen, A., Chen, C., & Chen, Y. (2010). Spoken emotion recognition using local fisher discriminant analysis. In *10th international conference on signal processing* (*ICSP*), *2010 IEEE* (pp. 538–540). Piscataway: IEEE.

Zhang, S., & Zhao, Z. (2008). Feature selection filtering methods for emotion recognition in Chinese speech signal. In *9th international conference on signal processing, 2008. ICSP 2008*. (pp. 1699–1702). Piscataway: IEEE.

Zheng, W. Q., Yu, J. S., & Zou, Y. X. (2015). An experimental study of speech emotion recognition based on deep convolutional neural networks. In *2015 international conference on affective computing and intelligent interaction* (*ACII*) (pp. 827–831). Piscataway: IEEE.

Zhou, J., Wang, G., Yang, Y., & Chen, P. (2006). Speech emotion recognition based on rough set and SVM. In *5th IEEE international conference on cognitive informatics, 2006. ICCI 2006*. (Vol. 1, pp. 53–61). Piscataway: IEEE.

Zhou, Y., Sun, Y., Yang, L., & Yan, Y. (2009). Applying articulatory features to speech emotion recognition. In *international conference on research challenges in computer science, 2009. ICRCCS'09*. (pp. 73–76). Piscataway: IEEE.

Zhu, L., Chen, L., Zhao, D., Zhou, J., & Zhang, W. (2017). Emotion recognition from chinese speech for smart affective services using a combination of SVM and DBN. *Sensors, 17*(7), 1694.

Zong, Y., Zheng, W., Zhang, T., & Huang, X. (2016). Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression. *IEEE Signal Processing Letters, 23*(5), 585–589.