Telecommun Syst (2013) 52:1467–1478
DOI 10.1007/s11235-011-9624-z

Speech emotion recognition approaches in human computer interaction

S. Ramakrishnan · Ibrahiem M.M. El Emary

Published online: 2 September 2011

Abstract Speech Emotion Recognition (SER) represents one of the emerging fields in human-computer interaction. Quality of the human-computer interface that mimics human speech emotions relies heavily on the types of features used and also on the classifier employed for recognition. The main purpose of this paper is to present a wide range of features employed for speech emotion recognition and the acoustic characteristics of those features. Also in this paper, we analyze the performance in terms of some important parameters such as: precision, recall, F-measure and recognition rate of the features using two of the commonly used emotional speech databases namely Berlin emotional database and Danish emotional database. Emotional speech recognition is being applied in modern human-computer interfaces and the overview of 10 interesting applications is also presented in this paper to illustrate the importance of this technique.

Keywords Speech emotion · Human-computer interface · Pitch and emotion recognition

1 Introduction

Understanding emotions is essential in human social interactions. Studies suggest that only 10% of human life is completely unemotional. Although having been studied since the 1950's, the investigation of emotional cues has made considerable advances in the last years [1, 2]. This is mainly due to the new application developments with respect to human-machine, human-robot interfaces and multimedia retrieval. From the technological point of view the reasons for the renewed interests are also due to: technological progress in recording, storing, and processing audio and visual information; the development of non-intrusive sensors; the advent of wearable computers; the urge to enrich human-computer interface from point-and-click to sense-and-feel.

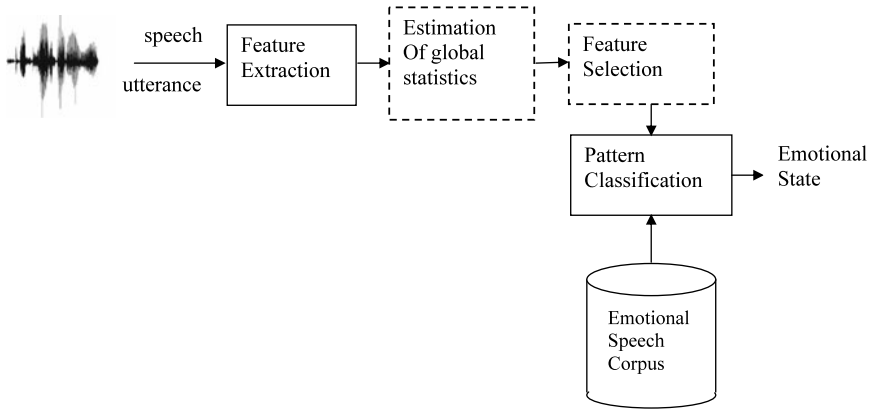
Emotion-oriented computing aims at the automatic recognition and synthesis of emotions in speech, facial expression, or any other biological channel [3–12]. Research about automated recognition of emotions in facial expressions is very rich [1, 13–16]. However, emotion recognition using facial recognition is computationally complex, because of that real-time implementation is prohibitive. Since it requires high quality cameras for capturing face images, cost of the implementation is also high. Besides human facial expressions, speech has been proven to be more promising modality for the recognition of human emotions. Vocal emotions are also an important constituent in multi-modal human computer interaction [17, 18]. In particular, speech emotion recognition is an important issue, as speech is the fundamental mode of human communication. Hence, in this article we focus on various aspects of speech emotion recognition (SER) methods.

The aim of SER is to enable a very natural interaction with the computer by speaking instead of using traditional input devices and not only have the machine understand the verbal content, but also more subtle cues such as affect that any human listener would easily react to. The interface between man and machine will become more meaningful if the machines can recognize the emotional contents. In this area, three different facets can be considered: (1) speech recog-

S. Ramakrishnan (✉)
Information Tech. Dep., Dr. Mahalingam College of Eng. & Tech., Udumalai Road, Pollachi 642003, India
e-mail: ram_f77@yahoo.com

I.M.M. El Emary
Faculty of Information Technology, King Abdulaziz University, P.O. Box 18388, Jeddah, King Saudi Arabia
e-mail: omary57@hotmail.com

Fig. 1 Overview of SER



nition in the presence of emotional speech, (2) synthesis of emotional speech, and (3) emotion recognition. This article focuses on the 3rd aspect i.e. emotion recognition from speech utterances.

SER has in the last decade shifted from a side issue to a major topic in human computer interaction and speech processing. SER has potentially wide applications. For example, human computer interfaces could be made to respond differently according to the emotional state of the user. This could be especially important in situations where speech is the primary mode of interaction with the machine [13].

In this article, we focus on (1) frame work and databases used for SER; (2) Acoustics characteristics of typical emotions of SER; (3) Various acoustic features employed for recognition of emotions from speech and (4) Applications of emotion recognition. Several reviews on emotional speech analysis have already appeared. However, to the best of our knowledge none of the reviews provide a comprehensive and up-to-date account on full spectrum of the research done in the field of SER. A few reviews focus on only data sets [13, 18–24], and some reviews our focuses on feature extraction methods. But this article covers full spectrum of SER starting from acoustic characteristics up to applications of SER.

This article is organized as follows: introduction section provides need for emotion recognition and in particular its need with respect to speech emotion recognition in the context of human-computer interaction. Section 2 provides framework and the databases used for performance analysis. Features that are critical in deciding the efficiency of SER are covered in Sect. 3. To the best of our knowledge none of the earlier articles provided a comprehensive account on acoustics characteristics of emotional speech and that is provided in Sect. 4. Various applications of SER are presented in Sect. 5. Conclusions are drawn in Sect. 6.

2 Framework and data bases for emotional recognition

2.1 Framework

Typical SER setup is depicted in Fig. 1. Dotted lined boxes indicate that those operations are optional. Following steps are normally involved in SER (1) Compute high-level statistical information from prosodic features at the sentence-level such as mean, range, variance, maximum, and minimum of F0 and energy. These statistics are concatenated to create an aggregated feature vector [25, 26]; (2) Estimate of global statistics of the features extracted. The global statistics are useful in speech emotion recognition, because they are less sensitive to linguistic information. These global statistics will be called simply as features throughout the paper; (3) Use of feature selection methods such as forward or backward feature selection, sequential forward floating search, genetic algorithms, evolutionary algorithms, linear discriminant analysis, or principal component analysis to search for a smaller set of relevant features [25]; (4) Employ an appropriate classifier to evaluate the performance of the system in terms of classification accuracy [19, 27–32].

2.2 Databases

One of the major needs of pattern recognition tasks is the data sets. A record of emotional speech data collections is undoubtedly useful not only for psychological studies but also for researchers interested in speech emotion recognition. The accuracy SER should be validated on significant dataset. Unfortunately, it is quite difficult to collect labeled emotional speech samples. In recent years, there has been a considerable amount of work on the collection of emotional speech. An overview, particularly on emotional speech databases, can be found in [20].

Emotional speech database contains any one of the 3 speech categories namely natural, simulated and elicited speech. Natural speech is simply spontaneous speech where

Table 1 Summary of 2 popular Emotional DB used in performance analysis

S. No.	Database details	EMO (Berlin Emotional Database)	DES (Danish Emotional Speech Database)
1	Nature	Acted	Acted
2	Language	German	Danish
3	Types of emotions, #	Anger-127 Boredom-79 Disgust-38 Fear-55 Joy-64 Neutral-78 Sadness-53	Anger-85 Happiness-86 Neutral-85 Sadness-84 Surprise-84
4	#Arousal	Low-248 High-246	Low-169 High-250
5	#Valence	Negative-352 Positive-142	Negative-169 Positive-250
6	Duration	22 minutes with 16 kHz	28 minutes with 20 kHz
7	No. of subjects	Total-10 Male-5 Female-5	Total-4 Male-2 Female-2
8	URL	http://pascal.kgw.tu-berlin.de/emodb/docu/#download	http://universal.elra.info/product_info.php?products_id=78

all emotions are real. Simulated or acted speech is speech expressed in a professionally deliberated manner. Finally, elicited speech is speech in which the emotions are induced. The elicited speech is neither neutral nor simulated.

Different types of databases are suitable for different purposes. There are objections against the use of acted emotions. However acted emotions are quite adequate for testing data. It is suitable for a novel method which first requires proof of the concept, rather than construction of a real-life application for the industry.

In this paper we consider two of the widely employed emotional speech databases for analyzing the performances of the features. An overview of the two emotional speech data collections is presented in Table 1. For each data collection, critical information such as speech language, the number and the profession of the subjects, male–female ratio, type and nature of emotions are presented.

3 Features for SER

To build an emotion recognition system, extraction of features that can truly represent the characteristics of the intended emotion is required. For emotional speech, a good reference model is the human hearing system. Several different types of prosody features have been explored in the literature. The goal is to simulate human perception of emotion, and identify possible features that can convey the underlying emotions in speech regardless of the language, speaker, and context.

The raw pitch, energy contours can be used as is, and are then called short-term features, or more often, the actual features are derived from these acoustic variables by applying (statistic) functions over the sequence of values within an emotion segment, thus called global statistics features. This could be e.g. mean pitch of a word or an utterance; further statistical measures are typically maximum, or minimum, etc. of the segment, but also regression, derivations or other more complex functions. The choice of feature type also determines the type of classifier. For global statistics features, a static classifier like Support Vector Machines (SVM), processing one instance at a time has to be used. Short-term features require a dynamic classifier such as Hidden Markov Models (HMM). One can say that in the first case, dynamic properties of emotions should be captured by the features, while in the latter case; they are dealt with by the classifier.

The features can be broadly classified into two categories namely acoustic and linguistic features. Among the two former is widely used and the classification of these features is illustrated in Fig. 2.

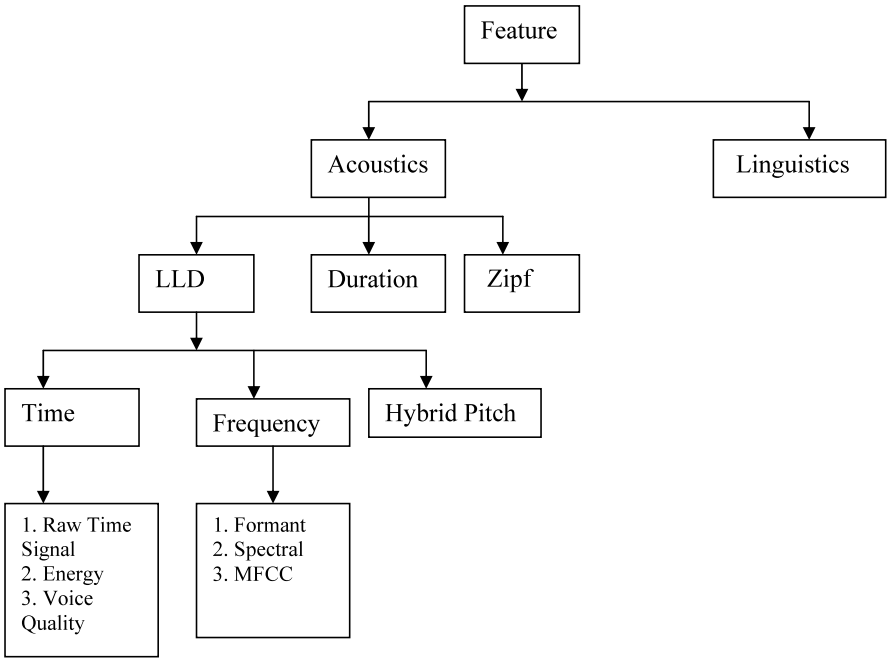


Fig. 2 Tree diagram for different types of features

Table 2 Types of Low-Level Descriptors (LLD) features

TYPE	LLD
Raw time signal	Elongation, Centroid, Zero-Crossing rule, Min, Max value, DC component
Energy	Log-energy and root mean square energy
Spectral	Energy in bands 0–250 HZ, 0–650 HZ, 250–650 HZ, 1000–4000 HZ 10%, 25%, 50%, 75%, and 90% Roll-off Centroid, Flux, and relative position of maximum and minimum
Pitch	Fundamental frequency F0 in Hz via Cepstrum and Autocorrelation (ACF)
Formants	F1–F7 frequency $+ \delta$, bandwidth $+ \delta$
MFCC & Cepstral	MFCC 0–12 and band 1–26
Voice quality	Harmonics to Noise Ratio (HNR), Probability of vocing and jitter

3.1 Low-Level Descriptors (LLD)

Commonly used Low-Level Descriptors (LLD) consists of spectral, raw time signal, formants, pitch, energy, MFCC and voice quality features [22, 32–34]. LLDs are derived from whole utterances. Thus, utterances of variable length can be mapped onto a feature vector of constant dimension. It is well known that different emotional states carry different prosodic patterns. Hence, prosodic feature like speech intensity, pitch and speaking rate can model prosodic patterns in different emotions. Similarly, spectral feature like MFCCs have been used successfully in emotion recognition. Cepstrum analysis is a source-filter\separation process commonly used in speech processing. Also voice quality features such as HNR, jitter, shimmer, spectral and cepstral features such as formants and MFCC have become more or less the "new standard". Table 2 provides comprehensive overview of the 7 LLD features.

3.2 Durational pause related features

The length and distribution of voiced and unvoiced segments is related to voice characteristics. The duration features fall out of the generative approach and they include the chunk length, measured in seconds, and the zero-crossing rate to roughly decode speaking rate. Furthermore, pause is obtained as the proportion of non-speech calculated by a voice activity detection algorithm from the signal energy and also approximated by the ratio of unvoiced pitch frames to the total number of pitch frames in the chunk [21, 35, 36].

Commonly used duration related features are: number of voiced and unvoiced regions; number of voiced and unvoiced frames; longest voiced and unvoiced region; ratio of number of voiced vs. unvoiced frames; ratio of number of voiced vs. unvoiced regions; ratio of number of voiced vs. total number of frames; ratio of number of voiced vs. total number of regions, jitter, and tremor.

3.3 Zipf features

Zipf features are used for a better rhythm and prosody characterization and also they prove to be very efficient in the valence dimension. The Zipf law is an empirical law proposed by G.K. Zipf. It says that the frequency of an event and its rank with respect to the frequency (from the most to the least frequent) are linked by a power law [30].

3.4 Linguistic features

Apart from acoustic features, also spoken or written text carries information about the underlying affective state. This is usually reflected in the usage of certain words or grammatical alterations. A number of approaches exist for this analysis: keyword spotting, rule-based modeling, semantic trees, latent semantic analysis, transformation-based learning, world-knowledge-modeling, key-phrase spotting, and Bayesian networks. Two methods seem to be predominant, presumably because they are shallow representations of linguistic knowledge and have already been frequently employed in automatic speech processing: (class-based) N-grams and vector space modeling [37–40].

4 Acoustic characteristics and performance analysis of the features used in SER

In this section, we review acoustic characteristics of various speech emotions such as anger, disgust, fear, joy, and sadness. Also analyze the performances of the features described in the previous section using two popular databases namely EMO DB and DES DB (presented in Sect. 2).

4.1 Acoustic characteristics of emotional speech

Speech emotion recognition can be done effectively only when acoustic characteristics of different emotions will carry distinct signatures. For instance, when one is in a state of anger, fear or joy, the sympathetic nervous system is aroused, the heart rate and blood pressure increase, the mouth becomes dry and there are occasional muscle tremors. Speech is then loud, fast and enunciated with strong high frequency energy. When one is bored or sad, the parasympathetic nervous system is aroused, the heart rate and blood pressure decrease and salivation increases, which results in slow, low-pitched speech with little high-frequency energy [2]. These kinds of acoustic characteristics will be exploited to form a unique signature for an emotion using the features discussed in the previous section.

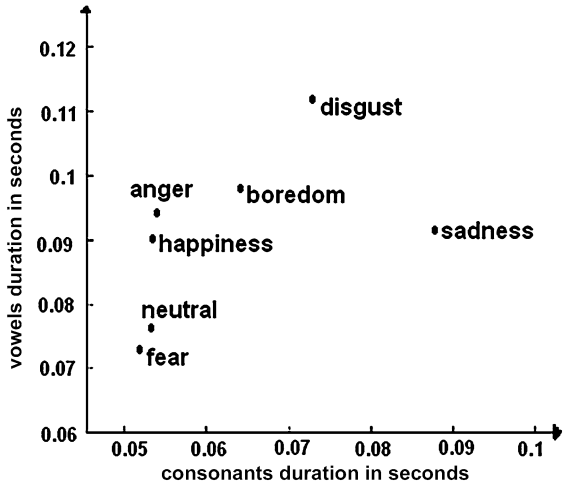


Fig. 3 Vowels and consonants duration from the Berlin database

Anger is the emotion of the highest energy and pitch level. It also has high variance and intensity range. Speaking rate of males is lower than that of females while expressing anger. In the sate of anger the rhythm is fast and syllables are accented but last word is not normally accented. Disgust is expressed with a low mean pitch and intensity levels. Speaking rate of disgust male is slower than that of females. The emotional state of fear is correlated with a high pitch level and a raised intensity level. The speaking rate of feared human is faster than disgusted human. When one is in the state of joy both mean pitch and intensity is high and speaking rate is also increased; few syllables and last word both are accented. Low levels of the mean intensity and mean pitch are measured when the subjects express sadness. The speech rate under similar circumstances is generally slower than that in the neutral state. The pitch contour trend is a valuable parameter, because it separates fear from joy. Fear resembles sadness having an almost downwards slope in the pitch contour, whereas joy exhibits a rising slope. The speech rate varies within each emotion. An interesting observation is that males speak faster when they are sad than when they are angry or disgusted [2].

Descriptors have been proposed to characterize the pitch and the energy for emotional data, a few can be found concerning the rhythm. A first attempt for rhythmic modelling can be obtained through the segmental durations [41]. To this end, both vowels and consonants duration were extracted from the transcription Berlin database. Results of the analysis are shown in Fig. 3. This figure illustrate that the emotions are reasonably well separated in the space.

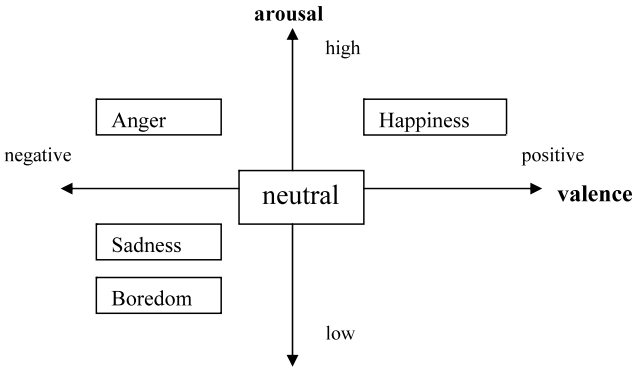


Fig. 4 Emotional space of the arousal valence

Psychological studies show prosody (pitch, intensity, and speaking rate) and voice quality to be most important to distinguish between emotions according to human perception. In particular pitch and intensity seem to be correlated to activation, so that high pitch and intensity values imply high, low pitch and intensity values low activation. Emotion states can be placed within a dimensional model of two or three affective dimensions. The dimensions are usually valence (from positive to negative) and arousal (from high to low), sometimes a third dimension like stance (from open to close) is added. In general, the arousal and valence dimensions can be used to distinguish the most basic emotions [42–44]. The emotion location in the arousal-valence space is shown in Fig. 4. Most emotional features are related to arousal. This suggests that the most emotional features confused anger and joy, and sadness and neutrality as valence-related emotions.

4.2 Performance analysis of the features used in SER

In order to analyze effectiveness of the features presented in Sect. 3 we conduct experiments on 2 datasets namely EMO (Berlin Emotional Dataset) and DES (Danish Emotional Speech Database) in terms of parameters such as recognition rate, precision, recall and F-measure. We have used software namely openSMILE and Praat for extraction of features and Matlab for performing the classification.

Figure 4 illustrates ability of arousal-valance dimension in discriminating the emotional state. In order to test the effectiveness of these 2 dimensions we have used pitch and MFCC features employing two different classifiers namely HMM and SVM. And the results are presented in Table 3. The Fig. 4 reveals that some emotional categories with high activation levels (i.e., high arousal) such as anger and happiness are clearly distinguished from neutral speech using pitch-related features. However, subdued emotional categories such as sadness present similar pitch characteristics to neutral speech. From Table 3, we can claim that the pitch and MFCC features are efficient in distinguishing between high-arousal emotions, e.g. anger, fear, and joy, versus low-arousal ones, e.g. sadness. Also these features effectively classify emotions which have similar arousal, e.g. Anger versus Joy. Between the classifiers SVM provides better recognition rate over HMM [27, 28, 45, 46].

Precision and recall are commonly used for evaluating the correctness of an emotional speech recognition algorithm. This two can complement the parameter, recognition rate, and can be treated as a simple metric that computes the fraction of instances for which the correct result is returned. Here, the set of possible labels for each emotional state is divided into two subsets, one of which is considered "relevant" for the purposes of the metric. Recall is then computed as the fraction of correct instances among all instances that actually belong to the relevant subset, while precision is the fraction of correct instances among those that the algorithm believes to belong to the relevant subset. Due to the slightly unbalanced class distribution, recognition is a rather less appropriate performance measure. Thus, we used the F-measure as the harmonic mean between recall and precision for performance evaluation. The experimental results are presented based on the 2 databases in Tables 4 and 5. An interesting result is that the precision rate is in general high, which means that there are not many neutral samples labeled as emotional (false positive).

In order to test the discrimination ability of the features presented in Sect. 3, experiments are carried out using SVM classifier on the 2 databases and the results are presented in Tables 6 and 7. Individual features are tested for its recognition efficiency and it can be seen from Tables 6 and 7 that, among various features pitch, MFCC and formants features are providing good recognition rates and hence a combination of these features are also used to test effectiveness.

Studies have shown that a hierarchical classification method achieves better performance than considering the features from all classes [23]; a hierarchical classification of emotions is shown in Fig. 5.

In general, humans have different vocal systems in terms of factors such as shape and size. This causes variations of emotional features from young age to middle and between male and female [46]. Error rate in detecting the emotions between young, middle-age & male and female is shown in Fig. 6. This result suggests that speaker-dependent emotional classification is providing better recognition over speaker-independent one. If hierarchical classification is also employed along with speaker-dependent system, one could achieve very high recognition rate than normally achieved.

5 Applications of SER in HCI

Speech Emotion Recognition (SER) has potentially wide applications in Human Computer Interaction (HCI). For ex-

Table 3 Recognition rate based on arousal and valence on the 2 popular DBs

S. No.	Classifier	EMO		DES	
		Arousal	Valence	Arousal	Valence
1	Pitch, formants and MFCC features with HMM	0.90	0.79	0.81	0.56
2	Pitch, formants and MFCC features with SVM	0.95	0.86	0.85	0.72

Table 4 Precision, recall, F-measure and recognition rate for various types of emotions in EMO DB

Type of emotions in EMO DB	Precision	Recall	F-measure	Recognition rate
Anger	0.86	0.83	0.84	0.96
Boredom	0.69	0.68	0.69	0.72
Disgust	0.61	0.59	0.60	0.71
Fear	0.67	0.62	0.64	0.81
Joy	0.84	0.85	0.85	0.95
Neutral	0.57	0.51	0.54	0.68
Sadness	0.79	0.71	0.75	0.75

Table 5 Precision, recall, F-measure and recognition rate for various types of emotions in DES DB

Type of emotions in DES DB	Precision	Recall	F-measure	Recognition rate
Anger	0.80	0.79	0.80	0.93
Happiness	0.78	0.77	0.78	0.81
Neutral	0.63	0.64	0.64	0.71
Sadness	0.81	0.76	0.78	0.74
Surprise	0.76	0.72	0.74	0.69

Table 6 Recognition rate versus various features for different emotions in EMO DB

Recognition rates for various emotions in EMO DB							
Type of features	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness
Raw time signal	0.80	0.67	0.64	0.69	0.81	0.45	0.63
Energy	0.87	0.68	0.65	0.70	0.90	0.48	0.67
Spectral	0.83	0.66	0.65	0.70	0.81	0.47	0.65
Pitch	0.91	0.70	0.67	0.74	0.90	0.59	0.68
Formants	0.92	0.69	0.68	0.79	0.92	0.62	0.70
MFCC & Cepstral	0.96	0.69	0.68	0.79	0.92	0.62	0.70
Voice quality	0.74	0.64	0.61	0.70	0.79	0.41	0.60
Durational pause related features	0.76	0.64	0.63	0.69	0.74	0.43	0.62
Zipf features	0.76	0.64	0.63	0.69	0.74	0.43	0.62
Combination of pitch, MFCC and formant features	0.96	0.72	0.71	0.81	0.96	0.68	0.75

ample, human computer interfaces could be made to respond differently according to the emotional state of the user. This could be especially important in situations where speech is the primary mode of interaction with the machine. Therefore, SER is an essential requirement to make the following applications more human-like and to increase their acceptance among potential users: (1) Robots, (2) Smart call centers, (3) Intelligent spoken tutoring systems, (4) Smart aircraft cockpits, (5) Prosody for dialog systems, (6) Ticket reservation system, (7) Diagnostic tool by therapists, (8) Information retrieval for medical analysis, (9) In-car board system, (10) Speech synthesis, (11) Computer games, (12) Sorting voice mail, (13) Intelligent toys, (14) Lie-detection, (15) Automatic searching in films and TV programs and (16) Telephone banking, etc.

Let us explain some of those.

Table 7 Recognition rate versus various features for different emotions in DES DB

Recognition rates for various emotions in DES DB					
Type of features	Anger	Happiness	Neutral	Sadness	Surprise
Raw time signal	0.82	0.81	0.60	0.64	0.60
Energy	0.87	0.85	0.61	0.70	0.61
Spectral	0.85	0.81	0.61	0.65	0.67
Pitch	0.89	0.87	0.64	0.73	0.67
Formants	0.91	0.88	0.68	0.75	0.67
MFCC & Cepstral	0.91	0.89	0.68	0.75	0.65
Voice quality	0.78	0.78	0.58	0.61	0.56
Durational pause related features	0.80	0.78	0.59	0.63	0.57
Zipf features	0.82	0.79	0.60	0.64	0.59
Combination of pitch, MFCC and formant features	0.93	0.80	0.71	0.74	0.70

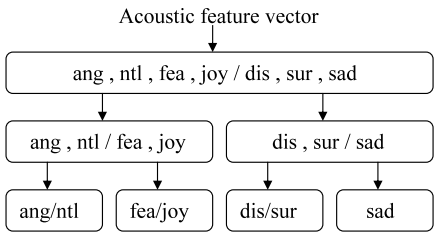


Fig. 5 Hierarchical classification of emotions in speech

5.1 Intelligent tutoring system

An Intelligent Tutoring System (ITS) replaces a human tutor by a machine. The system provides the student with a more personalized and friendly environment for learning according to their needs. The ITS is an educational software containing an artificial intelligence component and tracks the students work. This software infers strengths and weakness of the person and on that basis provides individual instructions. Imagine a system designed to teach piano. Such a system could detect if a student was overly frustrated, and provide them with simpler pieces to learn. Emotional speech recognition in tutoring system helps in improving the efficiency of knowledge transmission. The emotions are detected by means of lexical, prosodic, spectral, and syntactic analyses of users' speech. More frequently detected emotions in tutoring systems are hesitation, puzzle and confidence [2].

5.2 Lie detection

Lie detector using SER helps in deciding whether someone is lying or not. This mechanism is used particularly in areas such as Central Bureau of Investigation for finding out the criminals, cricket council to fight against corruption. X13-VSA PRO Voice Lie Detector 3.0.1 PRO is an innovative, advanced and sophisticated software system and a fully computerized voice stress analyzer that allows us to detect the truth instantly.

5.3 Telephone banking

Recently more and more banks are integrating emotional speech recognition into their Interactive Voice Response systems—Citibank, Wells Fargo and HSBC. Emotional speech recognition system in Telephone banking improves customer service levels. Recognizing emotions of customers helps in directing the customers to the specially trained personnel, rather than navigating through various levels of voice response systems to deal with their needs. EmoRate Software allows computers to detect emotions of the customers.

5.4 In-car board system

Emotional speech recognition in-car board system helps the human in performing various tasks such as steering, accelerating, braking, speed, route, distance to other vehicles, dimming, operating windscreen wipers, changing gears, and blinking. The ability of a car to understand natural speech provides a human-like driver assistance system. Emotional factors are decisive for enhanced safety and comfort while driving a car. Emotions such as aggressiveness, anger, confusion, sadness, sleepiness have major impact in causing road accidents. This can be minimized by recognizing drivers emotions and responding to those emotions by adapting to the current situation [2].

5.5 Prosody in dialog system

Prosody means the rhythm, melody and tone of the speech. Prosody can improve the interaction in the spoken dialog system by detecting the emotions, disfluencies and distinguishing statements from questions. Current dialog systems

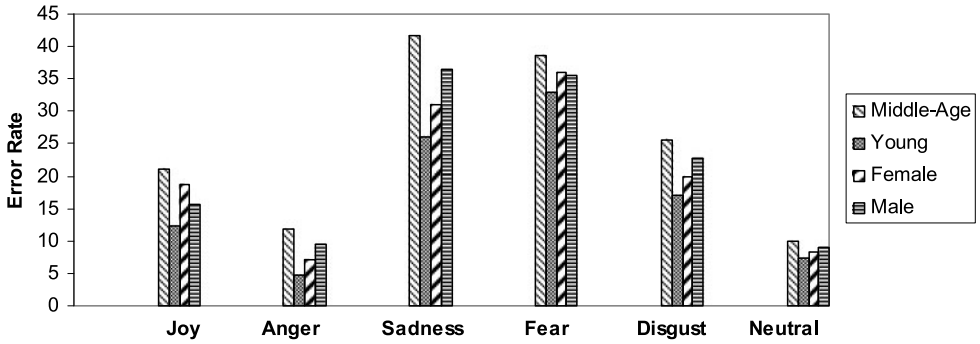


Fig. 6 Error rates for speaker-dependent emotion recognition system

often model prosody on the output side to generate acceptable speech synthesis, but few systems use prosody on the input side. To make the interaction in the dialog system smoother timing of response is considered. The prosodic and linguistic features extracted from the human speech has more effect on timing response [40].

5.6 Emotion recognition in call center

Emotional Speech Recognition is applied in call centers to detect the emotional state of the customers. Determining the emotional state helps the service provider to deal more effectively with the situation. Most frequently detected emotions are "agitation" which includes anger,happiness and fear, and "calm" which includes normal state and sadness. Depending on the emotion types,service providers will assign a proper agent to respond the message so as to benefit the company. Companies like Wisconsin Physician Services Insurance Corporation (WPSIC) have used emotion detection tool to save the fair share of their distressed policy holders by recognizing their caller's emotions and assigning proper agents to resolve the problem [2].

5.7 Sorting of voice mail

Voicemail is an electronic system for recording and storing of voice messages for later retrieval by the intended recipient. Recognition of emotions from speech is applied in sorting voice mail. Based on the emotions expressed by the caller voice mail messages are sorted [2].

5.8 Computer games

Computer games can be controlled through emotions of human speech. The computer recognizes human emotion from their speech and compute the level of game (easy, medium, hard). For example, if the human speech is in form of aggressive nature then the level becomes hard. Suppose if the human is too relaxed the level becomes easy. The rest of emotions come under medium level [2].

5.9 Diagnostic tool by speech therapists

Person who diagnosis and treats variety of speech, voice, and language disorders is called a Speech Therapist. By understanding and empathizing emotional stress and strains the therapists can know what the patient is suffering from. The software used for recording and analyzing the entire speech is icSpeech. The use of speech communication in healthcare is to allow the patient to describe their health condition to the best of their knowledge. In clinical analysis, human emotions are analysed based on features related to prosodics,the vocal tract, and parameters extracted directly from the glottal waveform. Emotional expressions can be referred by vocal affect extracted from the human speech [47].

5.10 Robots

Robots can interact with people and assist them in their daily routines, in common places such as homes, super markets, hospitals or offices. For accomplishing these tasks,robots should recognize the emotions of the humans to provide a friendly environment. Without recognizing the emotion, the robot cannot interact with the human in a natural way [48–50].

Authors would like to provide details such as software tools available for doing research in speech emotion recognition, projects already done internationally, conferences and systems based on emotional speech, etc. for the benefit of readers.

Emotional speech software tools: HMM Toolkit (HTK); openEAR; xwavesp package; FEELtrace; Praat; ANOVA; openSMILE: ESEDA feature extraction module; ESMERALDA; EmoVoice; EmoRate.

Emotional speech projects: VAESS; INTERFACE; PHYSTA; CREST-ESPproject; COCOSDA; CALLAS; Semaine; HUMAINE; FERMUS III; CEICES.

Systems which use emotions in speech: Sony AIBO Robot; Humanoids SDR3-X (Sony); SmartKom; SEMAINE; Jerk-O-Meter; Tiger's Furby.

Emotional Speech Conferences: InterSpeech Annual Conference; ISCA workshop; CEICES initiate.

6 Conclusions

Information on emotion is encoded in all aspects of language, in what we say and in how we say it, and the 'how' is even more important than the 'what'. This article focuses on analyzing the performances of the various features used in SER and also provides a comprehensive treatment on use of SER in human-computer interaction applications. The automatic recognition of emotion seems straight-forward. However, this is unfortunately not the case. First of all, psychological studies often get their insights from data of test persons acting to be in an emotional state. There, this clear mapping from acoustic variables might even be possible in a number of cases, though even when acting, intra- and inter-speaker variations are higher, as the expressivity of emotions is also dependent on the personality or the mood. In everyday human computer interaction, however, the occurring emotions are very spontaneous. There, these variations are considerably higher as these are not any more prototypical emotions but may be shaded, mixed, or weak and hardly distinguishable. This makes the task much harder, so that further acoustic features need to be investigated. Of course, personalized emotion recognition that is from only one speaker, is more reliable. Further evidence of the differences of acted and spontaneous emotions have been showed in human listening tests that the perception of acted emotions is different than that from natural emotions.

Technical advances in signal processing have allowed the development of a large number of speech technologies, many of which are on the brink of being applied and commercialized. However, so far real-time emotion recognition has scarcely been attempted and if so, only in prototypical applications, as there are still many problems that are not yet solved appropriately. A lot more signal processing modules are needed to close or at least narrow the gap between the human ability to observe and react to the affective state of the conversational partner and today's state-of-the-art human-computer interfaces.

References

1. Zeng, Z., Roisman, M. P. 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*(1), 39–58.
2. Vogt, T., Andre, E., & Wagner, J. (2008). Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realisation. In C. Peter & R. Beale (Eds.), *LNCS: Vol. 4868. Affect and emotion in HCI* (pp. 75–91).
3. Petrantonakis, P. C., & Hadjileontiadis, L. J. (2010). Emotion recognition from EEG using higher order crossings. *IEEE Transactions on Information Technology in Biomedicine*, *14*(2), 186–197.
4. Frantzidis, C. A., Bratsas, C., et al. (2010). On the classification of emotional biosignals evoked while viewing affective pictures: an integrated data-mining-based approach for healthcare applications. *IEEE Transactions on Information Technology in Biomedicine*, *14*(2), 309–318.
5. Lin, Y.-P., Wang, C.-H., Jung, T.-P., Wu, T.-L., Jeng, S.-K., Duann, J.-R., & Chen, J.-H. (2010). EEG-based emotion recognition in music listening. *IEEE Transactions on Biomedical Engineering*, *57*(7), 1798–1806.
6. Schaaff, K., & Schultz, T. (2009). Towards an EEG-based emotion recognizer for humanoid robots. In *The 18th IEEE international symposium on robot and human interactive communication*, Toyama, Japan, Sept. 27–Oct. 2 (pp. 719–722). University of Karlsruhe (TH), Karlsruhe, Germany.
7. Murugappan, M., Rizon, M., Nagarajan, R., Yaacob, S., Zunaidi, I., & Hazry, D. (2007). EEG feature extraction for classifying emotions using FCM and FKM. *International Journal of Computers and Communications*, *2*(1), 21–25.
8. Petrantonakis, P. C., & Hadjileontiadis, L. J. (2010). Emotion recognition from EEG using higher order crossings. *IEEE Transactions on Information Technology in Biomedicine*, *14*(2), 186–197.
9. Schaaff, K., & Schultz, T. (2009). Towards an EEG-based emotion recognizer for humanoid robots. In *The 18th IEEE international symposium on robot and human interactive communication*, Toyama, Japan, Sept. 27–Oct. 2 (pp. 792–796).
10. Lin, Y.-P., Wang, C.-H., Jung, T.-P., Wu, T.-L., Jeng, S.-K., Duann, J.-R., & Chen, J.-H. (2010). EEG-based emotion recognition in music listening. *IEEE Transactions on Biomedical Engineering*, *57*(7), 1798–1806.
11. International Conference on Information Technology and Computer Science (2009). The Research on Emotion recognition from ECG signal. In *International conference on information technology and computer science*, Kiev, July 25–26
12. Han, M.-J., Hsu, J.-H., & Song, K.-T. (2008). A new information fusion method for bimodal robotic emotion recognition. *Journal of Computers*, *3*(7), 39–47.
13. Chibelushi, C. C., Deravi, F., & Mason, J. S. D. (2002). A review of speech-based bimodal recognition. *IEEE Transactions on Multimedia*, *4*(1), 23–37.
14. Elwakdy, M., Elsehely, E., Eltokhy, M., & Elhennawy, A. (2008). Speech recognition using a wavelet transform to establish fuzzy inference system through subtractive clustering and neural network (ANFIS). *International Journal of Circuits, Systems and Signal Processing*, *4*(2), 264–273.
15. Ranjan, S. (2010). Exploring the discrete wavelet transform as a tool for Hindi speech recognition. *International Journal of Computer Theory and Engineering*, *2*(4), 642–645.
16. Grimm, M., Kroschel, K., & Narayanan, S. (2008). The Vera Am Mittag German audio-visual emotional Speech Database. In *IEEE international conference on multimedia & expo*, Hannover, Germany, 23–26 June.
17. Wollmer, M., Metallinou, A., Eyben, F., Schuller, B., & Narayanan, S. (2010). Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. In *International speech communication association*, Makuhari, Chiba, Japan, 26–30 September.
18. Firoz Shah, A., Raji Sukumar, A., & Babu Anto, P. (2010). Discrete wavelet transforms and artificial neural networks for speech emotion recognition. *International Journal of Computer Theory and Engineering*, *2*(3), 319–322.

19. Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., & Wendemuth, A. (2009). Acoustic emotion recognition: a benchmark comparison of performances. In *IEEE workshop on automatic speech recognition and understanding*, Merano, Italy, 13–20 December (pp. 552–557).
20. Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. (2003). Emotional speech: towards a new generation of databases. *Speech Communication*, *40*, 33–60.
21. Hansen, J. H. L. (1996). Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. *Speech Communication*, *20*(1–2), 151–170.
22. 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.
23. Xiao, Z., Dellandrea, E., Dou, W., Chen, L., & Ecole Centrale de Lyon (2007). Automatic hierarchical classification of emotional speech. In *Ninth IEEE international symposium on multimedia 2007—workshops* (pp. 291–296).
24. Camelin, N., Bechet, F., Damnati, G., & De Mori, R. (2010). Detection and interpretation of opinion expressions in spoken surveys. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(2), 369–381.
25. Ververidis, D., & Kotropoulos, C. (2008). Fast and accurate sequential floating forward feature selection with the Bayes classifier applied to speech emotion recognition. *Signal Processing*, *88*(12), 2956–2970.
26. Visser, E., Otsuka, M., & Lee, T.-W. (2003). A spatio-temporal speech enhancement scheme for robust speech recognition in noisy environments. *Speech Communication*, *41*, 393–407.
27. Nogueiras, A., Moreno, A., Bonafonte, A., & Mariño, J. B. (2001). Speech emotion recognition using hidden Markov models. In *EUROSPEECH 2001 Scandinavia, 7th European conference on speech communication and technology, 2nd INTERSPEECH Event*, Aalborg, Denmark, 3–7 September.
28. Neiberg, D., & Elenius, K. (2008). Automatic recognition of anger in spontaneous speech. In *INTERSPEECH 9th annual conference of the international speech communication association*, Brisbane, Australia, 22–26 September.
29. Schuller, B., Batliner, A., Steidl, S., & Seppi, D. (2009). Emotion recognition from speech: putting ASR in the loop. In *IEEE international conference on acoustics, speech, and signal processing (ICASSP 2009)*, Taipei, Taiwan, 19–24 April.
30. Xiao, Z., Dellandrea, E., & Chen, L. (2009). Recognition of emotions in speech by a hierarchical approach. In *Affective computing and intelligent interaction and workshops 2009 (ACII 2009), 3rd international conference*, Amsterdam, 10–12 September.
31. Khanchandani, K. B., & Hussain, M. A. (2009). Emotion recognition using multilayer perceptron and generalized feed forward neural network. *IEEE Journal of Scientific and Industrial Research*, *68*, 367–371.
32. Sobol-Shikler, T., & Robinson, P. (2010). Classification of complex information: inference of co-occurring affective states from their expressions in speech. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(7), 1284–1297.
33. Trimmer, C. G., & Cuddy, L. L. (2008). Emotional intelligence, not music training, predicts recognition of emotional speech prosody. *Emotion*, *8*(6), 838–849. Copyright 2008 by the American Psychological Association.
34. Erro, D., Navas, E. Hernáez, I., & Saratxaga, I. (2010). Emotion conversion based on prosodic unit selection. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(5), 974–983.
35. Ververidis, D., Kotropoulos, C., & Pitas, I. (2004). Automatic emotional speech classification. In *International speech communication association, acoustics, speech, and signal processing. Proceedings (ICASSP'04), IEEE international conference*, Quebec, Canada, 17–21 May.
36. Schuller, B., Seppi, D., Batliner, A., Maier, A., & Steidl, S. (2007). Towards more reality in the recognition of emotional speech. In *IEEE international conference on acoustics, speech, and signal processing*, (ICASSP), Honolulu, 15 April.
37. Truong, K. P., & Raaijmakers, S. (2008). Automatic recognition of spontaneous emotions in speech using acoustic and lexical features. In *LNCS: Vol. 5237. MLMI 2008* (pp. 161–172).
38. Schuller, B., Rigoll, G., & Lang, M. (2004). Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine—belief network architecture. In *IEEE international conference on acoustics, speech, and signal processing*, Quebec, Canada, 17–21 May.
39. Schuller, B., Vlasenko, B., Arsic, D., Rigoll, G., & Wendemuth, A. (2008). Combining speech recognition and acoustic word emotion models for robust text-independent emotion recognition. In *IEEE international conference on multimedia & expo*, Hannover, Germany, 23–26 June.
40. Fujie, S. Yagi, D., Matsusaka, Y., Kikuchi, H., & Kobayashi, T. (2004). Spoken dialogue system using prosody as para-linguistic information. In *Indian science congress association archive, speech prosody 2004, international conference*, Nara, Japan, 23–26 March.
41. Ringeval, F., & Chetouani, M. (2008). A vowel based approach for acted emotion recognition. In *INTERSPEECH 2008 9th annual conference of the international speech communication association*, Brisbane, Australia, 22–26 September.
42. 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.
43. Wöllmer, M., Schuller, B., Eyben, F., & Rigoll, G. (2010). Combining long short-term memory and dynamic Bayesian networks for incremental emotion-sensitive artificial listening. *IEEE Journal of Selected Topics in Signal Processing*, *4*(5), 867–881.
44. Tarng, W., Chen, Y.-Y., Li, C.-L., Hsie, K.-R., & Chen, M. (2010). Applications of support vector machines on smart phone systems for emotional speech recognition. *World Academy of Science, Engineering and Technology*, *72*, 106–113.
45. Chavhan, Y., Dhore, M. L., & Yesaware, P. (2010). Speech emotion recognition using support vector machine. *International Journal of Computer Applications*, *1*(20), 6–9.
46. Paulmann, S., Pell, M. D., & Kotz, S. A. (2008). How aging affects the recognition of emotional speech. *Brain and Language*, *104*, 262–269.
47. Moore, E., II, Clements, M. A., Peifer, J. W., Weisser, L. (2008). Critical analysis of the impact of glottal features in the classification of clinical depression in speech. *IEEE Transactions on Biomedical Engineering*, *55*(1), 96–107.
48. Jang, K.-D., & Kwon, O.-W. (2006). Speech emotion recognition for affective human-robot interaction. In *SPECOM'2006*, St. Petersburg, 25–29 June (pp. 419–422).
49. Park, J.-S., Kim, J.-H., & Oh, Y.-H. (2009). Feature vector classification based speech emotion recognition for service robots. *IEEE Transactions on Consumer Electronics*, *55*(3), 1590–1596.
50. Wang, Y., & Guan, L. (2008). Recognizing human emotional state from audiovisual signals. *IEEE Transactions on Multimedia*, *10*(4), 659–668.

S. Ramakrishnan received the B.E. degree in Electronics and Communication Engineering in 1998 from the Bharathidasan University, Trichy, and the M.E. degree in Communication Systems in 2000 from the Madurai Kamaraj University, Madurai. He received his Ph.D. degree in Information and Communication Engineering from Anna University, Chennai in 2007.

He has 10 years of teaching experience and 1 year industry experience. He is a Professor and the Head of the Department of Information Technology, Dr. Mahalingam College of Engineering and Technology, Pollachi, India.

Dr. Ramakrishnan is a Reviewer of 13 International Journals such as IEEE Transactions on Image Processing, IET Communications ACM Reviewer for Computing Reviews, Elsevier Science, International Journal of Vibration and Control, IET Generation, Transmission & Distribution, etc. He is in the editorial board of 4 International Journals. He is a Guest Editor of special issues in 2 international journals. He has published 45 papers in international, national journals and conference proceedings. Dr. S. Ramakrishnan has published a book for LAP, Germany. He has also reviewed 2 books for McGraw Hill International Edition and 1 book for ACM Computing Reviews. He is guiding 6 Ph.D. research scholars. His areas of research include digital image processing, soft computing, and digital signal processing.

Ibrahiem M.M. El Emary received the Dr.Eng. Degree in 1998 from the Electronic and Communication Department, Faculty of Engineering, Ain shams University, Egypt. Currently, he is a visiting Professor at King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia. His research interests cover: various analytic and discrete event simulation techniques, performance evaluation of communication networks, application of intelligent techniques in managing computer communication network, and performing a comparative studies between various policies and strategies of routing, congestion control, subnetting of computer communication networks. He published more than 150 articles in various refereed international journals and conferences covering: Computer Networks, Artificial Intelligent, Expert Systems, Software Agents, Information Retrieval, E-learning, Case Based Reasoning, Image Processing and Pattern Recognition and Robotic engineering. Also, in the current time, he is too interested in making a lot of scientific research in wireless sensor networks in view point of enhancing its algorithms of congestion control and routing protocols. Also, he participates in publishing five book chapters in two international books (published by IGI publisher and Springer Verlag) as well as co-editor of two books edited by two international publishers LAP Lampert. He joined MIR Lap in USA as a representative of KSA in this international research laboratory of Machine learning. He joins more than ten international refereed journals as an editor in chief, editor and reviewers. Finally, he joins also more than fifteen international refereed conferences as a membership in the technical and international committees.