Circuits, Systems, and Signal Processing (2020) 39:4459–4481
https://doi.org/10.1007/s00034-020-01377-y

Excitation Features of Speech for Emotion Recognition Using Neutral Speech as Reference

Sudarsana Reddy Kadiri[1] · P. Gangamohan[2] · Suryakanth V. Gangashetty[3] · Paavo Alku[1] · B. Yegnanarayana[3]

Received: 12 March 2019 / Revised: 11 February 2020 / Accepted: 14 February 2020 /
Published online: 25 February 2020

Abstract

In generation of emotional speech, there are deviations in the speech production features when compared to neutral (non-emotional) speech. The objective of this study is to capture the deviations in features related to the excitation component of speech and to develop a system for automatic recognition of emotions based on these deviations. The emotions considered in this study are anger, happiness, sadness and neutral state. The study shows that there are useful features in the deviations of the excitation features, which can be exploited to develop an emotion recognition system. The excitation features used in this study are the instantaneous fundamental frequency (F_0), the strength of excitation, the energy of excitation and the ratio of the high-frequency to low-frequency band energy (β). A hierarchical binary decision tree approach is used to develop an emotion recognition system with neutral speech as reference. The recognition experiments showed that the excitation features are comparable or better than the existing prosody features and spectral features, such as mel-frequency cepstral coefficients, perceptual linear predictive coefficients and modulation spectral features.

Keywords Excitation features · Emotion recognition · Zero frequency filtering (ZFF) · Linear prediction (LP) analysis · Short-time Fourier transform (STFT) · Kullback–Leibler (KL) distance

1 Introduction

The goal of speech technology is to make human–machine interaction as natural as possible. The two important modules of speech technology are automatic speech recognition and text-to-speech synthesis. The naturalness of interaction depends on the ability of the system to recognize and synthesize emotions in speech.

✉ Sudarsana Reddy Kadiri
sudarsana.kadiri@aalto.fi

Extended author information available on the last page of the article

Table 1 Trend in prosody features in emotional speech compared to neutral speech

	F_0 mean	F_0 variance	Energy	Speaking rate
Anger	inc	inc	inc	inc
Happiness	inc	inc	inc	inc
Sadness	dec	inc	dec	dec

inc increase, *dec* decrease

In recent years, efforts have been made in the field of emotion recognition. From the literature, it is observed that there are many interrelated issues such as databases, features, approaches and evaluation procedures that need to be considered for the development of an emotion recognition system. Ideally, databases consisting of natural 'spontaneous' emotions should be used in analysis of vocal emotions. However, it is difficult to collect such speech data due to privacy and copyright issues. Therefore, different research groups have collected several databases of emotional speech that can be categorized as simulated, seminatural and (near to) natural [12,29,47]. The simulated emotion corpus is recorded from professional speakers (actors) by prompting them to enact emotions through specified text in a given language. There are many examples of simulated databases such as the Berlin Emotional Speech Database (EMO-DB) [6] and the Danish Emotional Speech Database (DES) [14]. The seminatural database is also a kind of enacted data, where the context is given to the speakers. Examples of databases in this category are the USC-IEMOCAP corpus [30] (in English), and the German and Russian databases described in [12,29,47]. The third type of emotional speech databases is (near to) natural database, where recordings do not involve any prompting or the obvious eliciting of emotional responses. Sources for such natural situations are mostly from talk shows in TV broadcasts, interviews, group interactions, etc. [20]. The important aspects in collecting emotional databases and the description of the various types of databases are discussed in [12,63].

The set of features used for emotion recognition can be broadly characterized as prosodic and spectral features. The trend of the prosody features (including fundamental frequency (F_0), energy and speaking rate) in three emotion categories (anger, happiness and sadness) with respect to neutral state is given in Table 1 [37,46]. Similarly, the trend of the spectral features (including changes in formant frequencies and spectral tilt) is given in Table 2 [37,46]. There are some interconnections between the choice of features and the type of the database. For example, the deviations in spectral features such as formant frequencies and spectral tilt are analyzed in simulated parallel corpora. This is because the deviations in formant frequencies and spectral tilt can be compared only when the utterances of different emotion categories are of the same lexical content [37,43,46,59].

The existing emotion recognition approaches are motivated from applications such as speech recognition, speaker recognition and language identification [23,28,35]. In most of the studies [17,28,31,33,46,47], vectors consisting of spectral features like mel-frequency cepstral coefficients (MFCCs) and linear prediction cepstral coefficients (LPCCs), prosody features, energy features and their statistics are extracted from overlapping/non-overlapping segments of speech. For example, a large number of features are extracted in the open-source toolkit called OpenEAR [24,47,49,64]. Emotions

Table 2 Trend in spectral features in emotional speech compared to neutral speech

	F_1 mean	F_1 bandwidth	F_2 mean	Spectral tilt
Anger	inc	–	inc/dec	dec
Happiness	dec	inc	–	dec
Sadness	inc	dec	dec	inc

inc increase, *dec* decrease

are modeled using discriminative/non-discriminative models such as Gaussian mixture models (GMMs), auto-associative neural networks (AANNs), multilayer feedforward neural networks (MLFFNNs) and deep neural networks (DNNs) [28,29,32,51]. Binary classification techniques such as Bayesian logistic regression (BLR) and support vector machines (SVMs) are also used to classify the multi-class problem by adopting the hierarchical binary decision tree framework [23,24,30]. In [30], the authors used a binary decision tree approach using the features of the OpenSmile toolbox [48] (with a 384-dimensional feature set). In that study, neutral state was first distinguished from three emotions (anger, happiness and sadness), then in the next stage, sadness was distinguished from anger, and finally happiness was distinguished from sadness in the final stage. Recently, raw speech signals were used with deep neural networks for emotion recognition in [45,55].

It is to be noted that the performance in terms of recognition accuracy of emotion recognition systems using simulated parallel databases is high when compared to the systems using seminatural and natural databases [19,30,47,60]. This is because utterances of the same lexical content are used for training and testing, and on limited number of speakers data. As per analysis reported in [37,46], for speech segments of the same lexical content, there are deviations in the spectral features such as formants and spectral tilt. These deviations might help in the discrimination of emotions in the case of simulated parallel corpora.

A cross-corpora study was reported in [10,49], where cross tests were performed between corpora. The training of emotion models was developed using one corpus, and these models were tested with another corpus. The two corpora consisted of real-life call center data in French. The accuracy in percentage in the cross-corpora evaluation was reported to be 47% for three emotions (anger, neutral state and positive valence) [10]. Similarly in [49], intra-corpus and inter-corpus recognition of emotions was studied, and it was shown that the recognition accuracy depends on the specific group of emotions and feature combinations considered [40]. Most of the recent studies take advantage of various sets of features in recognition of emotions using sophisticated classification mechanisms [17,23,64].

The present study proposes a set of excitation features that are independent of language and lexical content. An approach for emotion recognition is proposed by characterizing emotions as deviations from neutral state. The objective is to analyze and capture these deviations using features related to the excitation component of the speech production system. The paper is organized as follows: In Sect. 2, background and motivation for exploring excitation features are discussed. Section 3 describes the emotional speech databases used in this study, and the extraction of the excitation

features. Analysis of the excitation features is given in Sect. 4. In Sect. 5, the proposed emotion recognition system is discussed. Experimental results are discussed in Sect. 6. Finally, Sect. 7 provides a summary of the work and a scope for further studies.

2 Background and Motivation for Exploring Excitation Features

In studying emotion recognition, it is necessary to process the speech signal suitably to capture emotion-specific information. Since emotional speech is produced by the human speech production mechanism, emotions can be analyzed using both the excitation (voice source) parameters and the vocal tract system parameters. In the literature, emotion recognition systems have been mostly studied using features representing vocal tract system characteristics. Only a few studies have analyzed emotional speech using voice source features [1,29,41,53,54,56,57]. Most of these studies [1,52–54,57] have focused mainly on specific utterances like vowels. For extraction of these voice source features, glottal flow estimates have been computed in these studies by using iterative adaptive inverse filtering (IAIF) [2].

In [56,57], the role of the voice source was analyzed in the perception of valence (positive and negative) and arousal (active and passive) from short vowels (150 ms), and it was shown that the normalized amplitude quotient (NAQ) correlates better with arousal than with valence for both genders. In [56,57], it was observed that in the vowels [i:] and [u:], the equivalent sound level was the only statistically significant variable in emotional expressions for synthetic data. Similarly, emotions in short segments of the vowel [a:] extracted from continuous speech were analyzed in [1], and it was shown that NAQ yielded significant differences for most of the emotions studied. Even though NAQ correlates with emotions, it has to be noted that NAQ by itself is not sufficient to discriminate different emotions accurately [1]. The interdependencies among the voice source features in emotional speech was studied for the sustained vowel [a:] in five emotions using six voice source parameters extracted from the glottal flow in [54]. In [52,53], robustness of glottal source features was studied in a cross-database scenario using four emotions (anger, happiness, sadness and neutral state).

Most of the studies that utilize voice source features in the analysis or recognition of emotions use glottal inverse filtering (GIF) to estimate the glottal flow from specific type of utterances, like vowels. Ideally, it would be preferable to derive the excitation features from the speech signal directly. However, it has been observed in many studies (e.g., [2,13,58]) that the performance of GIF deteriorates in high-pitched speech like in utterances produced by female or child speakers, and in emotional speech of high arousal. In addition, GIF might not work as well in continuous speech as in sustained vowel utterances and the performance of GIF is also affected when processing degraded speech [2,13,58]. Hence, it is justified to derive excitation features from the speech signal directly for the analysis and recognition of emotions.

2.1 Relation to Prior Work

In [21,22,36,39,62], attempts were made to derive some of the excitation features directly from the speech signal without computing the source-filter decomposition.

In [38,39,62], excitation features such as epochs/glottal closure instants, the strength of glottal closure and the instantaneous fundamental frequency were derived using the zero frequency filtering (ZFF) method [39]. In [21,22,36], the loudness feature was derived to capture the sharpness of glottal closure. To measure the changes in the closed to open phase regions of the glottis, a ratio between the high-frequency and the low-frequency spectral energies was proposed in [36]. In [15,16,27,41,42], some of these excitation features were used to study emotions in speech. In [16,27], the authors analyzed excitation features [instantaneous fundamental frequency (F_0), strength of excitation (SoE), energy of excitation (EoE) [16] and loudness (η)], which are extracted at the sub-segmental level of speech, for four emotions (anger, happiness, sadness and neutral state). The SoE parameter and the ratio of spectral energies between the high-frequency and the low-frequency ranges were used for discriminating angry and happy speech in [15]. In [41,42], features such as F_0 and SoE, and their first and second derivatives were used for the analysis and discrimination of emotions. In addition, in [18,41,44] the effect of emotions on the excitation of speech production was studied using prosody modification by converting speech in one emotion to another.

The present study is based on studying the relations among the parameters of the speech production mechanism using a physiologically motivated perspective. The study involves features of the excitation component of speech, namely the nature of the vocal fold vibration (at the glottis) [62], the strength of the impulse-like excitation at epoch [39], the energy around the epoch [21,22] and changes in the spectral features caused by the excitation such as the low-frequency spectral energy (LFSE) and the high-frequency spectral energy (HFSE) [36]. All these features are extracted from the speech signal directly without using GIF to estimate the glottal flow waveform like in [1,54,56,57]. An approach for emotion recognition is proposed by characterizing emotions as deviations of features from neutral speech.

In [21,22,36,38,39,62], methods were developed to derive excitation features from the speech signal. In [15,16,18,27,41,41,42,44], the authors analyzed excitation features [instantaneous fundamental frequency (F_0), strength of excitation (SoE), energy of excitation (EoE) and loudness (η)], which are extracted at the sub-segmental level of speech, for four emotions (anger, happiness, sadness and neutral state). Motivated by the good results achieved in [15,16,18,27,41,42,44], where it was shown that excitation features capture significant information about emotions, we make a systematic investigation in the analysis of features for two databases (including cases where lexical content is both same and different) and develop an emotion recognition system using neutral speech features as reference. This recognition system is developed based on the observation made in the feature analysis part of the study, which highlighted deviations in emotional features in a two-dimensional feature space compared to the two-dimensional feature space of neutral speech.

More specifically, the present study is an extension to the preliminary investigation in emotion recognition published in [27]. The extensions are as follows:

- A systematic analysis of the excitation features is carried out in four emotions (anger, happiness, sadness and neutral state).
- An emotion recognition system framework is developed using neutral speech as reference.

– The proposed emotion recognition system processes speech in short segments (2 s) and is therefore possible to be used in real-time applications.
– The excitation features studied are shown to be independent of lexical content.
– The effectiveness of the excitation features is also investigated in a cross-language scenario, where the system is trained using a database of one language and tested using a database of another language.

3 Emotional Speech Databases and Feature Extraction

Two types of emotional speech databases (seminatural and simulated) are used in this study. The description of the databases and the feature extraction procedure are discussed in Sects. 3.1 and 3.2, respectively.

3.1 Databases

3.1.1 The IIIT-H Telugu Emotional Speech Database

The IIIT-H Telugu Emotional Speech Database [16] is a seminatural database consisting of speech in the Indian language of Telugu collected from students of IIIT-Hyderabad. The data were collected from seven speakers (two females and five males) producing speech in four emotions (anger, happiness, sadness and neutral state). The students were asked to script a text themselves which helped them to generate emotional speech by remembering past situations and memories. All the recordings were carried out in laboratory environment using a close-microphone and electroglottography (EGG). For each emotion and for each speaker, the lexical content is different. The recordings were carried out in 2–3 sessions for each speaker, and the entire data consist of around 200 utterances. The database was evaluated in a perceptual listening test by 10 listeners for recognizability of the emotions. A total of 130 utterances were used in the current study, and they consisted of 35, 27, 34 and 34 utterances in anger, happiness, neutral state and sadness, respectively. The mean duration of each utterance is approximately 3 s.

3.1.2 The Berlin Emotional Database (EMO-DB)

The Berlin Emotional Speech Database (EMO-DB) [6] is a German database that was recorded in an anechoic chamber at Technical University of Berlin. Ten (five males and five females) professional native actors were asked to speak 10 sentences in seven emotions (anger, happiness, neutral state, sadness, fear, disgust and boredom) in one or more sessions. The entire data set consists of around 800 utterances. The database was evaluated in a perception test with 20 listeners regarding the recognizability of emotions. The utterances were selected that had recognition rate better than 80% and naturalness better than 60%. The mean duration of each utterance is approximately 3 s. In this study, utterances in four emotions (anger, happiness, neutral state and sadness)

are considered, and the number of utterances in each emotion is 127, 71, 79 and 62, respectively.

3.2 Extraction of Excitation Features

Motivated by the studies in [16], excitation features are used to develop an emotion recognition system. The excitation features used consist of the following parameters: the instantaneous fundamental frequency (F_0) [62], the strength of excitation (SoE) [39], the energy of excitation (EoE) [21,22] and the ratio between the high-frequency and the low-frequency spectral energies (β) [36]. These features are extracted using the zero frequency filtering (ZFF) method [38,62], linear prediction (LP) analysis [34] and short-time Fourier transform (STFT) [3].

The glottal closure instants (GCIs) of speech are obtained using the ZFF method [39]. In this method, the speech signal is passed through a cascade of two ideal digital resonators located at 0 Hz, followed by trend removal. The resultant signal is called the ZFF signal. The negative-to-positive zero crossings of the ZFF signal correspond to the GCIs. The interval between two successive GCIs gives the fundamental period T_0. The instantaneous fundamental frequency is given by $F_0 = 1/T_0$. The slope of the ZFF signal at each GCI is called the strength of excitation (SoE), which is related to the amplitude of the impulse-like excitation in most cases [39]. As the ZFF signal exhibits high energy in the voiced regions, the energy of the ZFF is used to detect voiced and unvoiced regions [11].

Linear prediction (LP) residual gives an approximation of the excitation component of the speech signal [34]. The energy of excitation (EoE) parameter is computed from the samples of the Hilbert envelope of the LP residual over a 2-ms region around each GCI. This gives a measure of vocal effort [21,22]. A 10th-order LP analysis is used for each 16-ms frame using a 2-ms frame shift.

A segmental feature, which is the ratio between the high-frequency and low-frequency spectral energy (β), was proposed in [36] for discriminating shouting and neutral speech. It was shown that β is related to the effects caused by the changes in the vocal fold vibration characteristics between the two styles of vocalization. As the β feature captures arousal characteristic of speech, the feature is expected to be useful also in emotion recognition. The β measure is computed as the ratio of the high-frequency band (800–4000 Hz) energy to the low-frequency band (0–550 Hz) energy from short-time Fourier magnitude spectrum of speech signal.

4 Analysis of Excitation Features

The impulse-like excitation produced by the abrupt closure of the vocal folds is an important characteristic of the speech excitation [61]. Moreover, temporal regions around GCIs correspond to regions of high SNR in the speech signal. Hence, in this study, we focus on the features extracted in these regions of high SNR in the speech signal. The features (F_0, SoE, EoE and β) are computed from speech using the ZFF method [38,62], LP analysis [34] and short-time Fourier transform [3]. The mean and

Table 3 Mean and standard deviation (SD) of the excitation parameters for emotional speech in the IIIT-H database

	F_0 (in Hz)		SoE		EoE		β		F_1 (in Hz)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Speaker 1										
Neutral state	186	32	14	6.4	0.3	0.31	3.0	0.55	512	115
Anger	291	48	2.2	0.8	2.7	2.12	4.2	0.47	556	132
Happiness	278	37	2.7	2.5	1.3	1.13	3.6	0.76	514	137
Sadness	172	27	16	5.4	0.3	0.33	3.1	0.61	523	120
Speaker 2										
Neutral state	155	18	16	7.1	0.3	0.24	2.5	0.7	482	145
Anger	225	37	4.4	1.8	3	2.64	6.2	0.8	555	143
Happiness	205	29	5.8	3.8	1	0.93	3.9	0.9	475	145
Sadness	140	23	21	9.3	0.2	0.15	2.6	0.75	485	150

Table 4 Mean and standard deviation (SD) of the excitation parameters for emotional speech in the EMO-DB database

	F_0 (in Hz)		SoE		EoE		β		F_1 (in Hz)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Speaker 1										
Neutral state	193	31	9.4	4.8	0.4	0.34	3.5	0.76	516	124
Anger	294	39	1.2	0.4	2.3	1.92	5.8	0.59	545	119
Happiness	265	37	2.4	1.2	1.1	0.93	3.9	0.71	524	127
Sadness	196	33	10	4.6	0.4	0.21	3.3	0.66	516	123
Speaker 2										
Neutral state	130	16	20	6.3	0.5	0.42	3.1	0.63	471	133
Anger	231	34	8.1	1.9	2.7	1.78	5.9	0.53	512	135
Happiness	196	30	10	4.2	1.3	0.87	3.6	0.64	489	129
Sadness	119	19	23	6.1	0.4	0.25	2.2	0.58	492	140

standard deviation of the distributions of the excitation features for two speakers (one female and one male) using five utterances for each emotion from the IIIT-H Telugu and German EMO-DB databases are given in Tables 3 and 4.

From Tables 3 and 4, it is observed that F_0 in anger and happiness is high compared to neutral state [16,41]. But comparatively, happiness shows a slightly lower value of the average F_0 than anger. For sadness, the average F_0 is mostly lower than that in neutral speech. This is in line with previous studies published in [4,7,43,59].

It is interesting to note that the strength of the impulse-like excitation (SoE) in anger appears to be lower than that in neutral speech. This is due to the decrease in the length of the pitch period (T_0) in anger. In order to maintain the high rate of vibration, the vocal folds may not close with high suction, which results in lower values of SoE in anger. For the same reason, happiness shows a lower SoE, but this parameter is

Table 5 Characteristics of the excitation features for emotional speech with respect to neutral speech

	F_0	SoE	EoE	β	F_1
Anger	High	Very low	Very high	High	High
Happiness	High	Low	High	High	Low
Sadness	Low	High	Low	Low	High

still higher than in anger. The variance of SoE is very low in anger when compared to happiness, even though the mean values of SoE are similar [41]. In the case of sadness, SoE is higher due to the large periodicity (T_0) associated with it.

The EoE parameter is computed from the samples of the Hilbert envelope of the LP residual over a 2-ms region around each GCI. The EoE is higher in anger than in happiness and lower in sadness compared to neutral speech [16]. Note that this is different from the energy computed from the speech signal directly. Hence, the energy of the excitation component is a better indicator of vocal effort.

The spectral band energy ratio (β) is related to the effects of changes in the vocal fold vibration characteristics, and it captures loudness or arousal characteristics of speech [36]. It can be observed that β is high in anger and happiness and low in sadness. The reason for the high β values in anger and happiness is that the high-frequency band energy is large due to a longer glottal closed phase and vice versa in sadness. The mean and standard deviation values of the first formant frequency (F_1) are also given in Tables 3 and 4. The mean of F_1 (400–750 Hz) is slightly larger in anger compared to neutral state, but there is no clear difference between happiness and sadness compared to neutral speech. The standard deviations of β and F_1 are similar in all emotions.

The above observations are with respect to neutral speech of the speaker. From Tables 3 and 4, it is important to note that the dynamic ranges of the features are speaker specific. For example, F_0 in neutral speech of male speaker 2 is similar to that of speech in sadness by female speaker 1. But for a given speaker, the main trends in the feature values are emotion specific.

Thus, the analysis results of the excitation features in emotional speech with respect to neutral speech can be summarized as shown in Table 5. The excitation features show discrimination among the emotions even though there exists some correlation between anger and happiness.

In order to capture the relations among the features, two features are considered at a time to form a two-dimensional (2-D) feature space. As F_0, SoE, EoE are extracted around GCIs, they are considered in pairs, and the segmental features β and F_1 are used as another pair. Hence, four 2-D feature spaces ($C1$ to $C4$) are formed as follows:

$C1$: (F_0 vs SoE),
$C2$: (EoE vs F_0),
$C3$: (EoE vs SoE), and
$C4$: (β vs F_1).

To analyze the emotion-specific deviations in these 2-D feature spaces, the reference (neutral) and test (emotional) utterances of the same speaker are considered together. For each reference utterance (neutral state), four 2-D feature spaces (corresponding to anger, happiness, sadness and neutral state) are obtained. As an illustration, Fig. 1

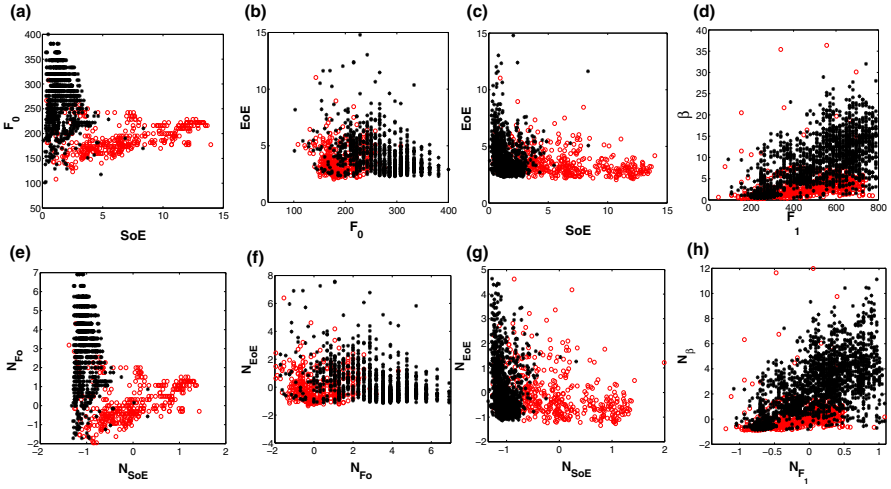

Fig. 1 Distribution for four combinations of the 2-D feature pairs between a male speaker's reference (neutral) utterance (marked by 'o') and emotional (anger) utterance (marked by '*'). Figures 1(a), 1(b), 1(c) and 1(d) are computed before the normalization, and Figs. 1(e), 1(f), 1(g) and 1(h) after the normalization (where 'N' refers to normalization)

shows the 2-D distributions for a reference utterance (neutral state, indicated by 'o') and test utterance (anger, indicated by '*') of the same speaker. The deviations in feature spaces between anger and neutral state can be observed from the figure. For example, in the feature space $C1$ (F_0 vs SoE) in Fig. 1a, F_0 increases and SoE decreases in anger. Similarly, the changes in the feature spaces for all emotions can be observed with respect to neutral speech.

From the analysis of the excitation features of emotional speech (given in Tables 3, 4), it is observed that the variance of the features shows discrimination even though the mean values of the features indicate less discrimination. Hence, it is useful to capture the divergence between features extracted from neutral and emotional speech signals. In order to utilize divergence, the Kullback–Leibler (KL) distance [9] is used. The distribution in the 2-D feature space of each utterance is modeled by a Gaussian probability distribution function, which is represented by mean vector and covariance matrix. The KL distance is computed between the corresponding 2-D feature distributions of the reference and test utterances as follows:

$$D_{\mathrm{KL}} = \frac{1}{2}\left(tr\left(\Sigma_1^{-1}\Sigma_0\right) + (\mu_1-\mu_0)^T\,\Sigma_1^{-1}\,(\mu_1-\mu_0)\right) - \frac{1}{2}\left(k + \ln\left(\frac{\det\Sigma_0}{\det\Sigma_1}\right)\right) \tag{1}$$

where D_{KL} is the KL distance, k is the dimension of the distribution, Σ_0, Σ_1 are the covariance matrices of the distributions of the feature pair of reference (neutral) and test (emotional) utterances, respectively, and μ_0, μ_1 are the corresponding mean vectors.

Table 6 The average KL distances between reference (neutral) utterances and test utterances in different emotions (anger, happiness, sadness and neutral state) of the IIIT-H database involving different lexical contents

	KL distances			
	C1	C2	C3	C4
Speaker 1				
Neutral versus neutral	0.6	0.4	0.5	0.3
Neutral versus anger	20	180	350	2.3
Neutral versus happiness	8	40	129	2.1
Neutral versus sadness	6	3	3	1.2
Speaker 2				
Neutral versus neutral	1.7	0.4	1.2	0.2
Neutral versus anger	7.2	50	60	2.1
Neutral versus happiness	5.6	35	45	1.3
Neutral versus sadness	6.2	3	3	0.9

Table 7 The average KL distances between reference (neutral) utterances and test utterances in different emotions (anger, happiness, sadness and neutral state) of the EMO-DB database involving same lexical contents

	KL distances			
	C1	C2	C3	C4
Speaker 1				
Neutral versus neutral	1.5	0.9	0.9	0.4
Neutral versus anger	40	80	240	1.7
Neutral versus happiness	6	50	50	1.2
Neutral versus sadness	5	1.8	8	0.9
Speaker 2				
Neutral versus neutral	0.5	0.3	1.2	0.9
Neutral versus anger	40	150	450	1.6
Neutral versus happiness	15	40	50	1.3
Neutral versus sadness	4	1.3	8	0.8

Using the 2-D feature space, the KL distances between the reference (neutral) utterance and all other emotions are shown in Table 6 for two speakers of the IIIT-H database. It can be clearly seen that the KL distances between the reference neutral utterances and the test neutral utterances are lower compared to the KL distances between the reference neutral utterance and the test emotional utterances in anger, happiness and sadness.

Results of the excitation feature analysis obtained for two speakers of the EMO-DB database are given in Tables 7 and 8. Table 7 corresponds to the case where the utterances in all emotions are of the same lexical contents, whereas the data in Table 8 correspond to having utterances of different lexical contents in the four emotion categories. Similar observations can be made as in the case of the IIIT-H database (Table 6). It appears that the characteristics of the excitation features are independent of the lexical content.

It is important to note that the KL distances vary between the speakers (i.e., variability due to speaker) and also between the emotions of the speakers (i.e., variability

Table 8 The average KL distances between reference (neutral) utterances and test utterances in different emotions (anger, happiness, sadness and neutral state) of the EMO-DB database involving different lexical contents

	KL distances			
	C1	C2	C3	C4
Speaker 1				
Neutral versus neutral	2.1	1.7	1.2	0.5
Neutral versus anger	76	210	293	3.2
Neutral versus happiness	42	133	180	1.4
Neutral versus sadness	9.8	1.8	23	0.7
Speaker 2				
Neutral versus neutral	0.4	1.1	1.1	0.3
Neutral versus anger	49	114	256	2.5
Neutral versus happiness	21	86	43	0.9
Neutral versus sadness	7.5	3.2	4	0.4

due to emotion). The speaker variability is mainly due to variations in dynamic ranges of feature values between speakers. From Tables 6, 7 and 8, it can also be observed that the KL distances of all feature combinations in the case of anger (test) utterances are high most of the time for both databases. This indicates that anger shows large deviations from neutral speech in both Telugu and German. When the test utterance corresponds to sadness, the KL distances for all four feature combinations are closer to neutral state, indicating that sadness may not deviate much from neutral state. This has also been observed in other studies in emotion recognition [28,47,48]. For developing emotion recognition system using these excitation source features, it is necessary to capture the speaker variability and emotion variability.

5 Emotion Recognition System based on Excitation Features

In order to capture the variability within the speakers, the distributions of neutral utterances are normalized as follows. Let us denote the values of F_0, SoE, EoE, β and F_1 for a reference neutral utterance by R_{F_0}, R_{SoE}, R_{EoE}, R_β and R_{F_1}, respectively, and for an emotional utterance by E_{F_0}, E_{SoE}, E_{EoE}, E_β and E_{F_1}, respectively. Let $R_{m_{F_0}}$, $R_{m_{\text{SoE}}}$, $R_{m_{\text{EoE}}}$, R_{m_β} and $R_{m_{F_1}}$, respectively, represent the mean values of the distributions of R_{F_0}, R_{SoE}, R_{EoE}, R_β and R_{F_1}. Likewise, let $R_{\sigma_{F_0}}$, $R_{\sigma_{\text{SoE}}}$, $R_{\sigma_{\text{EoE}}}$, R_{σ_β} and $R_{\sigma_{F_1}}$ represent the standard deviations of the distributions of R_{F_0}, R_{SoE}, R_{EoE}, R_β and R_{F_1}, respectively.

The distributions of neutral utterances are normalized with respect to mean and standard deviation as follows. The normalized distributions for R_{F_0} are given by:

$$N_{R_{F_0}} = \frac{R_{F_0} - R_{m_{F_0}}}{R_{\sigma_{F_0}}}. \tag{2}$$

Similarly, the values of the normalized distributions $N_{R_{\text{SoE}}}$, $N_{R_{\text{EoE}}}$, N_{R_β} and $N_{R_{F_1}}$ are obtained for R_{SoE}, R_{EoE}, R_β and R_{F_1}, respectively. The normalized distributions for

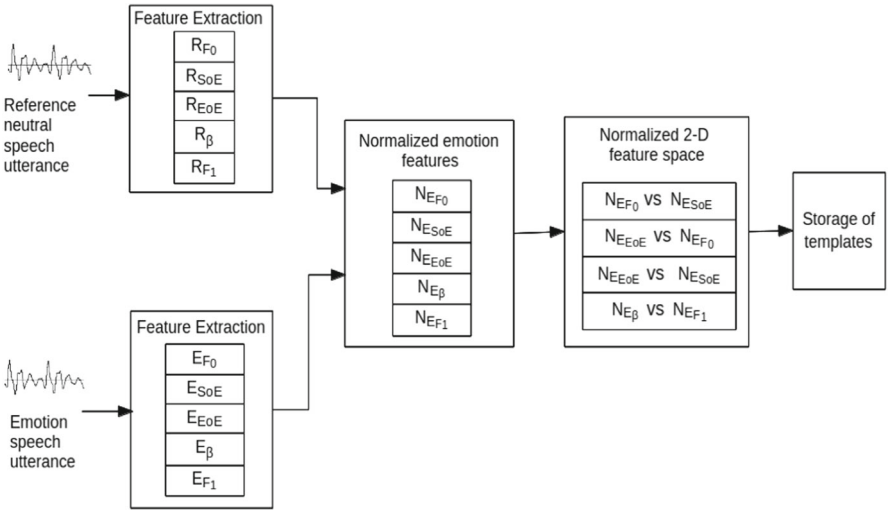

Fig. 2 Training phase (template generation process) of the emotion recognition system

the neutral utterance are shown by 'o' in Fig. 1e–h for the distributions of the neutral utterance in Fig. 1a–d, respectively.

To capture the variability due to emotions of a speaker, the distributions of features of an emotion utterance are normalized with respect to the neutral utterance as follows. The normalized distribution of E_{F_0} is given by:

$$N_{E_{F_0}} = \frac{E_{F_0} - R_{m_{F_0}}}{R_{\sigma_{F_0}}}. \tag{3}$$

Similarly, the values of the normalized distributions $N_{E_{\text{SoE}}}$, $N_{E_{\text{EoE}}}$, N_{E_β} and $N_{E_{F_1}}$ are obtained for E_{SoE}, E_{EoE}, E_β and E_{F_1}, respectively. The normalized distributions for the emotional (anger) utterance are shown by '*' in Fig. 1e–h for the distributions of the emotional (anger) utterance in Fig. 1a–d, respectively.

The normalization is done in a speaker-specific manner using the speaker's neutral utterance. This helps in reducing the variability among different speakers.

Four two-dimensional (2-D) feature distributions are formed by using the following combinations:

$D1$: ($N_{E_{F_0}}$ versus $N_{E_{\text{SoE}}}$),
$D2$: ($N_{E_{\text{EoE}}}$ versus $N_{E_{F_0}}$),
$D3$: ($N_{E_{\text{EoE}}}$ versus $N_{E_{\text{SoE}}}$), and
$D4$: (N_{E_β} versus $N_{E_{F_1}}$).

Each of these 2-D feature distributions is modeled by a Gaussian distribution, represented by mean vector and covariance matrix.

The training and testing phases of the proposed emotion recognition system are as shown in Figs. 2 and 3, respectively.

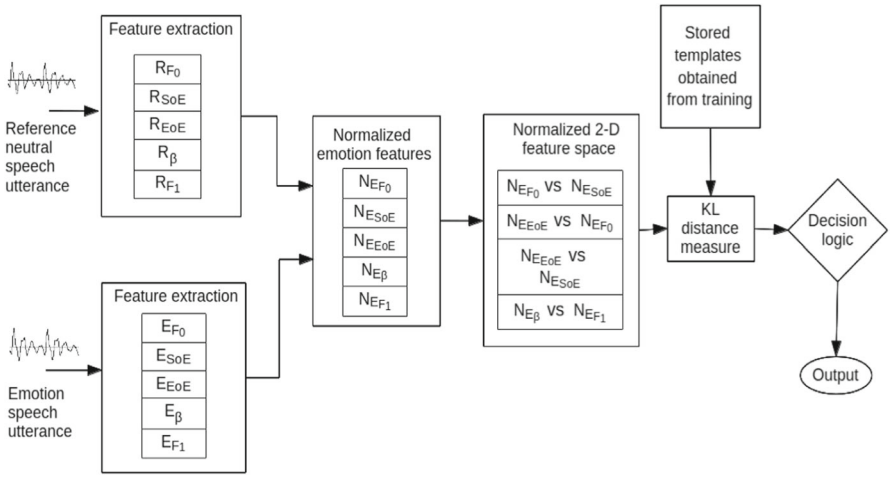

Fig. 3 Testing phase of the emotion recognition system

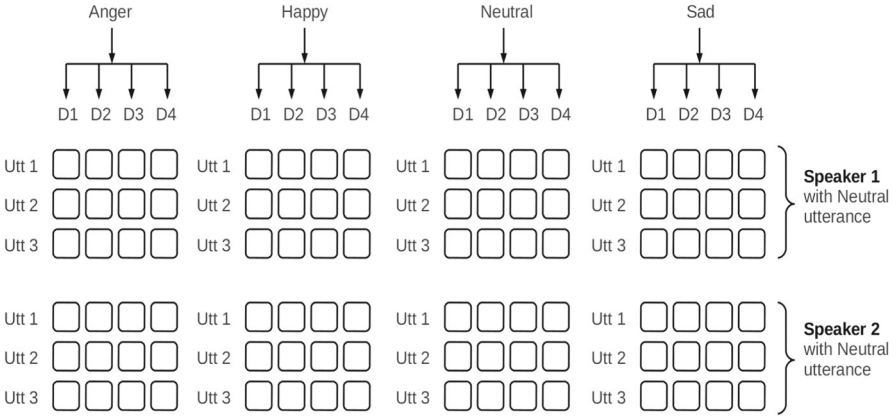

Fig. 4 An illustration of the stored templates for two speakers of the IIIT-H database. Here Utt1, Utt2 and Utt3 refer to the Utterance 1, Utterance 2 and Utterance 3 of the corresponding emotion, respectively, and D1, D2, D3 and D4 refer to the normalized 2-D emotion distribution

The training process involves generation of templates. Reference templates are generated by using three utterances for each of the four emotions (anger, happiness, sadness and neutral state) from seven speakers of the IIIT-H database. An additional neutral utterance from each of the seven speakers is also used. Therefore, a total of $(3 \times 4 \times 7 = 84; 84 + 7)$ 91 utterances are used. For each utterance (of the 84 utterances), four normalized distributions are generated using a neutral utterance of the corresponding speaker, and these distributions are called templates. Hence, $84 \times 4 = 336$ templates are created. The storage of mean vector and covariance matrix of the normalized 2-D emotion distribution is referred to as the stored template. As an illustration, a plan of the stored templates for two of the seven speakers from the IIIT-H database is shown in Fig. 4.

Table 9 Confusion matrix for emotions using the maximum number of output emotion labels for the IIIT-H database

	Neutral state	Sadness	Anger	Happiness
Neutral state	31/34	3/34	0/34	0/34
Sadness	6/34	24/34	1/34	3/34
Anger	2/35	1/35	28/35	4/35
Happiness	3/27	2/27	6/27	16/27

For testing, a neutral utterance and an emotional utterance are collected from the test speaker to derive the normalized emotion features. For each test case, the distributions of the features of the emotional utterance are normalized with respect to the neutral utterance, as in Fig. 1e–h. The normalized features of the test utterance (test templates) are compared with each of the corresponding normalized features of the trained templates using the KL distance.

With three utterances for each of the four emotions for each reference (trained) speaker, there are 12 utterances. For each utterance, there are four distributions (D1, D2, D3 and D4). Thus, for each test utterance, we get $12 \times 4 = 48$ KL distances for each reference (trained) speaker. The KL distances for each 2-D feature and emotion category are averaged over the three utterances, thus giving a total of $4 \times 4 = 16$ averaged KL distances per speaker, for the four pairs of the 2-D features and for the four emotion categories. The lowest of the averaged KL distances (for a given 2-D feature combination) across the four emotions is used to determine the emotion label for that 2-D feature. Thus, we get four emotion labels for each reference speaker. Leaving out the comparison with the test speaker in the reference, there are six other reference speakers providing $6 \times 4 = 24$ emotion labels for a given test utterance.

6 Results and Discussion

From the 24 emotion labels for each test utterance, the emotion with the maximum number of emotion labels is selected as the emotion category for the test utterance. The resulting confusion matrix is given in Table 9. Note that all the experiments are carried out with leave-one-speaker-out (LOSO) cross-validation.

From the results given in Table 9, it is observed that the confusion between anger and happiness is high. Similar observations are made between sadness and neutral state. This is because the features such as F_0 show an increasing trend and SoE shows a decreasing trend for anger and happiness when compared to neutral speech [16,41]. In the case of sadness, these excitation source features are not changing remarkably when compared to neutral state.

In order to improve the performance, a 2-stage binary decision logic [30] is implemented as shown in Fig. 5. In Stage 1, anger and happiness are grouped into one class, and sadness and neutral state are grouped into another class. The final decision of the emotion category is obtained from Stage 2, where comparisons are made between neutral state versus sadness, and between anger versus happiness using the following

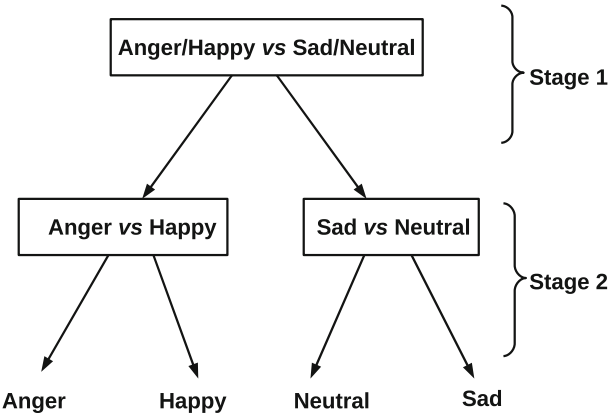

Fig. 5 Block diagram of binary tree decision logic [27,30]

Table 10 Confusion matrix after Stage 1 in binary tree decision logic for the IIIT-H database

	Neutral state/sadness	Anger/happiness
Neutral state/sadness	66/68	2/68
Anger/happiness	3/62	59/62

Table 11 Confusion matrix after Stage 2 in binary tree decision logic for the IIIT-H database

	Neutral state	Sadness	Anger	Happiness
Neutral state	32/34	2/34	0/34	0/34
Sadness	4/34	28/34	1/34	1/34
Anger	0/35	0/35	30/35	5/35
Happiness	1/27	1/27	7/27	18/27

decision criteria. Between neutral speech and sadness, neutral state is chosen, if the number of neutral labels $>$ (the number of sad labels $+$ 3). This is because there is high correlation between the features of neutral utterances compared to those of sad utterances. This is also evident from the KL distances of the feature combinations given in Tables 6, 7 and 8. Similarly, between anger and happiness, anger is chosen, if the number of anger labels $>$ (the number of happy labels $+$ 2).

The confusion matrices after Stage 1 and Stage 2 are given in Tables 10 and 11, respectively. From the results given in Table 10, the binary classification at stage 1 gives an accuracy of 96%. The number of original neutral/sad utterances recognized as angry/happy is reduced and vice versa. This is in line with the previous studies which have investigated acoustic features that are effective in discriminating emotions of high activation (anger, happiness) from emotions of low activation (sadness, boredom) [26, 30,50]. From Table 11, it is observed that the confusion between anger and happiness remains high, whereas the performance to discriminate sadness and neutral state has improved. The recognition for neutral state, sadness, anger and happiness is 94.1%,

Table 12 Confusion matrix after Stage 1 in binary tree decision logic for the EMO-DB database

	Neutral state/sadness	Anger/happiness
Neutral state/sadness	138/141	3/141
Anger/happiness	2/198	196/198

Table 13 Confusion matrix after Stage 2 in binary tree decision for the EMO-DB database

	Neutral state	Sadness	Anger	Happiness
Neutral state	74/79	3/79	0/79	2/79
Sadness	27/62	34/62	0/62	1/62
Anger	0/127	0/127	112/127	15/127
Happiness	2/71	0/71	32/71	37/71

Table 14 Emotion recognition results obtained for EMO-DB with the proposed method and with the SVM classifier using baseline feature sets based on spectral features (MFCC [60], MSF [60] and PLP [60]), prosody features [60] and the combination of the excitation features and MFCCs

	Proposed	MFCC	MSF	PLP	Prosody	Excitation + MFCCs
Recognition (%)	76	65.6	74	63.2	75	78.6

82.4%, 85.7% and 66.7%, respectively, giving an average recognition accuracy of 82.3% for the 4-class problem.

The proposed emotion recognition system was also evaluated using the EMO-DB database, and the results are given in Tables 12 and 13 after Stage 1 and Stage 2, respectively.

For the EMO-DB database, the recognition accuracy at Stage 1 is 98%. The recognition accuracy for the 4-class problem after Stage 2 is 76%. The performance of the system for the data of the EMO-DB database is low because of confusions between anger and happiness. The proposed excitation features were compared with the prosody features [60] and three short-term spectral features (mel-frequency cepstral coefficients (MFCCs) [60], perceptual linear predictive coefficients (PLPs) [25,60] and modulation spectral features (MSFs) [60]) using a SVM classifier [8] with leave-one-speaker-out (LOSO) cross-validation [60]. Table 14 shows the emotion recognition results obtained using the baseline feature sets (MFCCs, PLPs, MSFs and prosody features) [60] with the SVM classifier, the proposed emotion recognition results with the excitation features, and the combination of the proposed excitation features with the MFCCs using the SVM classifier with LOSO cross-validation. From Table 14, it can be observed that the results obtained using the excitation features are comparable or better than the existing prosody features and spectral features (MFCCs, PLPs and MSFs). Furthermore, it can be observed that there exists complimentary information between the proposed excitation features and the MFCC features. In [24], a large number of multiple feature sets (6552 features extracted using the openEAR toolkit) and various SVM schemes were used for a language-dependent and speaker-independent system

Table 15 Confusion matrix for the emotion classifier processing a 2-s speech buffer using the IIIT-H database

	Neutral state	Sadness	Anger	Happiness
Neutral state	112/130	12/130	2/130	4/130
Sadness	29/141	107/141	0/141	5/141
Anger	0/102	0/102	86/102	16/102
Happiness	2/71	0/71	12/71	57/71

Table 16 Confusion matrix of the emotion classifier processing a 2-s speech buffer using the EMO-DB database

	Neutral state	Sadness	Anger	Happiness
Neutral state	130/143	10/143	0/143	3/143
Sadness	66/212	141/212	0/212	5/212
Anger	0/267	0/267	224/267	43/267
Happiness	6/143	0/143	66/143	71/143

using the EMO-DB database and the study reported a recognition accuracy of 79.5%. It is to be noted that the focus of the present study is on the excitation features and their behavior in different emotions rather than unraveling which combination of feature toolkits and back-ends results in the best emotion recognition accuracy.

The proposed emotion recognition system is also used for online testing. For this, speech utterances of each speaker are concatenated with emotion labels intact. This is done for all the speakers in both of the databases. A neutral utterance of the corresponding speaker is used as a reference. The testing is carried out by processing a 2-s speech buffer of the test speech signal. The confusion scores for the online testing with the IIIT-H and EMO-DB databases are shown in Tables 15 and 16, respectively.

From the results given in Tables 15 and 16, the recognition accuracy of the IIIT-H and EMO-DB databases is 81.5% and 73.9%, respectively. Although there is some loss of suprasegmental information because of the 2-s speech buffering, there is not much reduction in performance. This is because the proposed features use only the subsegmental information around the epochs. One reason for the reduction in performance is that all the segments of an utterance may not show similar distributions of emotional information. This is also evident from [4,5,7], where it was shown that the emotionally salient aspects of speech are important in recognition and synthesis of emotional speech.

To test the effectiveness of the excitation features for cross-language databases, the reference templates created (i.e., training) using the EMO-DB database are used to test with data from the IIIT-H database and vice versa. From the results given in Tables 17 and 18, the recognition accuracy is about 68% in the former and 61% in the latter for the 4-class problem. This indicates that to some extent language and cultural aspects in expressing vocal emotions affect in recognition of emotions. However, it is

Table 17 Confusion matrix of emotion classifier for training with the German EMO-DB database and testing with the IIIT-H Telugu database

	Neutral state	Sadness	Anger	Happiness
Neutral state	107/130	18/130	0/130	5/130
Sadness	45/141	76/141	7/141	13/141
Anger	10/102	2/102	73/102	17/102
Happiness	6/71	2/71	18/71	45/71

Table 18 Confusion matrix of emotion classifier for training with the IIIT-H Telugu database and testing with the German EMO-DB database

	Neutral state	Sadness	Anger	Happiness
Neutral state	96/143	42/143	0/143	5/143
Sadness	63/212	136/212	4/212	9/212
Anger	46/267	45/267	158/267	18/267
Happiness	10/143	3/143	54/143	76/143

worth emphasizing that the extracted excitation features are independent of the lexical content.

The results of the proposed method indicate that the features corresponding to vocal effort seem to carry emotion-specific information. The performance of the system may be improved by increasing the number of reference (trained) templates and speakers. As there are confusions between anger and happiness, and between sadness and neutral state, deriving features which are more emotion specific may reduce the confusion between them.

7 Conclusions

In this paper, features corresponding to speech excitation were studied in analysis and recognition of vocal emotions. The emotion recognition system based on the features related to excitation component of speech production was developed by considering emotion states as deviations from neutral state. The deviations were captured through 2-D feature spaces. A template-based representation of 2-D normalized feature distributions of emotions using neutral speech as reference was generated from training examples. The emotion recognition system uses reference templates derived from the utterances in anger, happiness, sadness and neutral state. Although the system is speaker independent, a neutral speech utterance of the speaker is required for registration before testing. This is because of variability of dynamic ranges of the excitation features across speakers. One advantage of the proposed method is that it can be used in the recognition of emotions from short segments (2 s) of speech.

Ideally, an emotion recognition system should recognize the emotion category of speech without having access to neutral speech of the speaker. In this sense, the current study is limited. Existing emotion recognition systems have been developed using

mainly features representing the vocal tract system characteristics. Since the present study demonstrates that the excitation features capture effectively the emotion-specific characteristics of speech, it may be possible to combine the features from the excitation and the vocal tract system to improve the overall performance of emotion recognition systems. In addition, exploring the relations among the excitation features in an emotion-specific way might help for developing more robust emotion recognition systems.

Acknowledgements Open access funding provided by Aalto University. A part of this work was carried out when the first and second authors were at Speech Processing Laboratory, International Institute of Information Technology-Hyderabad, India. The first author would like to thank the Academy of Finland (Project 312490) for supporting his stay in Finland as a postdoctoral researcher. The last author would like to thank the Indian National Science Academy (INSA) for their support.

Compliance with ethical standards

Conflict of interest The authors declare no competing financial interests.

References

1. M. Airas, P. Alku, Emotions in vowel segments of continuous speech: analysis of the glottal flow using the normalized amplitude quotient. Phonetica **63**(1), 26–46 (2006)
2. P. Alku, Glottal inverse filtering analysis of human voice production-a review of estimation and parameterization methods of the glottal excitation and their applications. Sadhana **36**(5), 623–650 (2011)
3. J.B. Allen, L. Rabiner, A unified approach to short-time Fourier analysis and synthesis. Proc. IEEE **65**(11), 1558–1564 (1977)
4. J.P. Arias, C. Busso, N.B. Yoma, Shape-based modeling of the fundamental frequency contour for emotion detection in speech. Comput. Speech Lang. **28**(1), 278–294 (2014)
5. M. Bulut, S. Narayanan, On the robustness of overall F0-only modifications to the perception of emotions in speech. J. Acoust. Soc. Am. **123**(6), 4547–4558 (2008)
6. F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, B. Weiss, A database of German emotional speech, in *INTERSPEECH* (2005), pp. 1517–1520
7. C. Busso, S. Lee, S. Narayanan, Analysis of emotionally salient aspects of fundamental frequency for emotion detection. IEEE Trans. Audio Speech Lang. Process. **17**(4), 582–596 (2009)
8. C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, 2 July 2007
9. T.M. Cover, J.A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991)
10. L. Devillers, C. Vaudable, C. Chastagnol, Real-life emotion-related states detection in call centers: a cross-corpora study, in *INTERSPEECH* (2010), pp. 2350–2353
11. N. Dhananjaya, B. Yegnanarayana, Voiced/nonvoiced detection based on robustness of voiced epochs. IEEE Signal Process. Lett. **17**(3), 273–276 (2010)
12. E. Douglas-Cowie, N. Campbell, R. Cowie, P. Roach, Emotional speech: towards a new generation of databases. Speech Commun. **40**(1–2), 33–60 (2003)
13. T. Drugman, B. Bozkurt, T. Dutoit, A comparative study of glottal source estimation techniques. Comput. Speech Lang. **26**, 20–34 (2012)

14. I.S. Engberg, A. Varnich Hansen, O. Andersen, P. Dalsgaard, Design, recording and verification of a Danish emotional speech database, in *EUROSPEECH* (ISCA, 1997), pp. 1695–1698
15. P. Gangamohan, S.R. Kadiri, S.V. Gangashetty, B. Yegnanarayana, Excitation source features for discrimination of anger and happy emotions, in *INTERSPEECH* (2014), pp. 1253–1257
16. P. Gangamohan, S.R. Kadiri, B. Yegnanarayana, Analysis of emotional speech at subsegmental level, in *INTERSPEECH*, August (2013), pp. 1916–1920
17. M.J. Gangeh, P. Fewzee, A. Ghodsi, M.S. Kamel, F. Karray, Multiview supervised dictionary learning in speech emotion recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(6), 1056–1068 (2014)
18. D. Govind, S.R.M. Prasanna, B. Yegnanarayana, Neutral to target emotion conversion using source and suprasegmental information, in *Interspeech* (2011), pp. 2969–2972
19. M. Grimm, K. Kroschel, E. Mower, S. Narayanan, Primitives-based evaluation and estimation of emotions in speech. Speech Commun. **49**(10–11), 787–800 (2007)
20. M. Grimm, K. Kroschel, S.S. Narayanan, The Vera am Mittag German audio-visual emotional speech database, in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, Hannover, Germany, June (2008), pp. 865–868
21. S. Guruprasad, Significance of processing regions of high signal-to-noise ratio in speech signals. PhD Thesis, Apr (2011)
22. S. Guruprasad, B. Yegnanarayana, Performance of an event-based instantaneous fundamental frequency estimator for distant speech signals. IEEE Trans. Audio Speech Lang. Process. **19**(7), 1853–1864 (2011)
23. A. Hassan, R. Damper, M. Niranjan, On acoustic emotion recognition: compensating for covariate shift. IEEE Trans. Audio Speech Lang. Process. **21**(7), 1458–1468 (2013)
24. A. Hassan, R.I. Damper, Classification of emotional speech using 3DEC hierarchical classifier. Speech Commun. **54**(7), 903–916 (2012)
25. H. Hermansky, Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Am. **87**(4), 1738–1752 (1990)
26. J.H. Jeon, R. Xia, Y. Liu, Sentence level emotion recognition based on decisions from subsentence segments, in *ICASSP* (2011), pp. 4940–4943
27. S.R. Kadiri, P. Gangamohan, S.V. Gangashetty, B. Yegnanarayana, Analysis of excitation source features of speech for emotion recognition, in *INTERSPEECH* (2015), pp. 1324–1328
28. M. Kockmann, L. Burget, J. Cernocký, Application of speaker- and language identification state-of-the-art techniques for emotion recognition. Speech Commun. **53**(9–10), 1172–1185 (2011)
29. S.G. Koolagudi, K. Sreenivasa Rao, Emotion recognition from speech: a review. Int. J. Speech Technol. **15**(2), 99–117 (2012)
30. C.-C. Lee, E. Mower, C. Busso, S. Lee, S. Narayanan, Emotion recognition using a hierarchical binary decision tree approach. Speech Commun. **53**(9–10), 1162–1171 (2011)
31. C.M. Lee, S.S. Narayanan, Toward detecting emotions in spoken dialogs. IEEE Trans. Audio Speech Lang. Process. **13**(2), 293–303 (2005)
32. L. Li, Y. Zhao, D. Jiang, Y. Zhang, F. Wang, I. Gonzalez, E. Valentin, H. Sahli, Hybrid deep neural network–hidden Markov model (DNN–HMM) based speech emotion recognition, in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction* (2013), pp. 312–317
33. I. Luengo, E. Navas, I. Hernáez, Feature analysis and evaluation for automatic emotion identification in speech. IEEE Trans. Multimed. **12**(6), 490–501 (2010)
34. J. Makhoul, Linear prediction: a tutorial review. Proc. IEEE **63**, 561–580 (1975)
35. A. Milton, S. Tamil Selvi, Class-specific multiple classifiers scheme to recognize emotions from speech signals. Comput. Speech Lang. **28**(3), 727–742 (2014)
36. V.K. Mittal, B. Yegnanarayana, Effect of glottal dynamics in the production of shouted speech. J. Acoust. Soc. Am. **133**(5), 3050–3061 (2013)
37. D. Morrison, R. Wang, L.C. De Silva, Ensemble methods for spoken emotion recognition in call-centres. Speech Commun. **49**(2), 98–112 (2007)
38. K.S.R. Murty, B. Yegnanarayana, Epoch extraction from speech signals. IEEE Trans. Audio Speech Lang. Process. **16**(8), 1602–1613 (2008)
39. K.S.R. Murty, B. Yegnanarayana, M. Anand Joseph, Characterization of glottal activity from speech signals. IEEE Signal Process. Lett. **16**(6), 469–472 (2009)
40. T. Pfister, P. Robinson, Real-time recognition of affective states from nonverbal features of speech and its application for public speaking skill analysis. IEEE Trans. Affect. Comput. **2**(2), 66–78 (2011)

41. S.R.M. Prasanna, D. Govind, Analysis of excitation source information in emotional speech, in *INTERSPEECH*. ISCA (2010), pp. 781–784
42. D. Pravena, D. Govind, Significance of incorporating excitation source parameters for improved emotion recognition from speech and electroglottographic signals. Int. J. Speech Technol. **20**(4), 787–797 (2017)
43. K.R. Scherer, R. Banse, Acoustic profiles in vocal emotion expression. J. Personal. Soc. Psychol. **70**(3), 614–636 (1996)
44. K. Sreenivasa Rao, S.G. Koolagudi, Characterization and recognition of emotions from speech using excitation source information. Int. J. Speech Technol. **16**, 181–201 (2013)
45. M. Sarma, P. Ghahremani, D. Povey, N.K. Goel, K.K. Sarma, N. Dehak, Emotion identification from raw speech signals using DNNs, in *INTERSPEECH* (2018), pp. 3097–3101
46. K.R. Scherer, Vocal communication of emotion: a review of research paradigms. Speech Commun. **40**(1–2), 227–256 (2003)
47. B. Schuller, A. Batliner, S. Steidl, D. Seppi, Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. Speech Commun. **53**(9–10), 1062–1087 (2011)
48. B. Schuller, S. Steidl, A. Batliner, The interspeech 2009 emotion challenge, in *INTERSPEECH* (2009), pp. 312–315
49. B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, G. Rigoll, Cross-corpus acoustic emotion recognition: variances and strategies. IEEE Trans. Affect. Comput. **1**(2), 119–131 (2010)
50. M. Shami, W. Verhelst, An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. Speech Commun. **49**(3), 201–212 (2007)
51. A. Stuhlsatz, C. Meyer, F. Eyben, T. ZieIke, G. Meier, B. Schuller, Deep neural networks for acoustic emotion recognition: raising the benchmarks, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2011), pp. 5688–5691
52. R. Sun, E. Moore, A preliminary study on cross-databases emotion recognition using the glottal features in speech, in *INTERSPEECH* (2012), pp. 1628–1631
53. R. Sun, E. Moore, J.F. Torres, Investigating glottal parameters for differentiating emotional categories with similar prosodics, in *ICASSP* (2009), pp. 4509–4512
54. J. Sundberg, S. Patel, E. Björkner, K.R. Scherer, Interdependencies among voice source parameters in emotional speech. IEEE Trans. Affect. Comput. **2**(3), 162–174 (2011)
55. P. Tzirakis, J. Zhang, B.W. Schuller, End-to-end speech emotion recognition using deep neural networks, in *ICASSP* (2018), pp. 5089–5093
56. T. Waaramaa, A.-M. Laukkanen, M. Airas, P. Alku, Perception of emotional valences and activity levels from vowel segments of continuous speech. J. Voice **24**(1), 30–38 (2010)
57. T. Waaramaa-Mäki-Kulmala, T. Yliopisto, *Emotions in Voice: Acoustic and Perceptual Analysis of Voice Quality in the Vocal Expression of Emotions* (Acta univesitatis Tamperensis. Tampere University Press, Tampere, 2009)
58. J. Walker, P. Murphy, A review of glottal waveform analysis, in *Progress in Nonlinear Speech Processing. Lecture Notes in Computer Science*, vol. 4391, ed. by Y. Stylianou, M. Faundez-Zanuy, A. Esposito (Springer, Berlin, 2007), pp. 1–21
59. C.E. Williams, K.N. Stevens, Emotions and speech: some acoustical correlates. J. Acoust. Soc. Am. **52**(4B), 1238–1250 (1972)
60. S. Wu, T.H. Falk, W.-Y. Chan, Automatic speech emotion recognition using modulation spectral features. Speech Commun. **53**(5), 768–785 (2011)
61. B. Yegnanarayana, S.V. Gangashetty, Epoch-based analysis of speech signals. Sadhana **36**(5), 651–697 (2011)
62. B. Yegnanarayana, K. Sri Rama Murty, Event-based instantaneous fundamental frequency estimation from speech signals. IEEE Trans. Audio Speech Lang. Process. **17**(4), 614–624 (2009)
63. Z. Zeng, M. Pantic, G.I. Roisman, T.S. Huang, A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. **31**(1), 39–58 (2009)
64. W. Zheng, M. Xin, X. Wang, B. Wang, A novel speech emotion recognition method via incomplete sparse least square regression. IEEE Signal Process. Lett. **21**(5), 569–572 (2014)

Affiliations

Sudarsana Reddy Kadiri[1] · P. Gangamohan[2] · Suryakanth V. Gangashetty[3] · Paavo Alku[1] · B. Yegnanarayana[3]

P. Gangamohan
ganga39@klh.edu.in

Suryakanth V. Gangashetty
svg@iiit.ac.in

Paavo Alku
paavo.alku@aalto.fi

B. Yegnanarayana
yegna@iiit.ac.in

1 Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland

2 Koneru Lakshmaiah Education Foundation (KL University), Hyderabad, India

3 Speech Processing Laboratory, International Institute of Information Technology-Hyderabad, Hyderabad, India