IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 5, 2020, accepted March 20, 2020, date of publication March 24, 2020, date of current version April 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2982954

Excitation Features of Speech for Speaker-Specific Emotion Detection

SUDARSANA REDDY KADIRI, (Member, IEEE), AND PAAVO ALKU, (Fellow, IEEE)
Department of Signal Processing and Acoustics, Aalto University, FI-00076 Espoo, Finland

Corresponding author: Sudarsana Reddy Kadiri (sudarsana.kadiri@aalto.fi)

This work was supported in part by the Academy of Finland under Project 312490.

ABSTRACT In this article, we study emotion detection from speech in a speaker-specific scenario. By parameterizing the excitation component of voiced speech, the study explores deviations between emotional speech (e.g., speech produced in anger, happiness, sadness, etc.) and neutral speech (i.e., non-emotional) to develop an automatic emotion detection system. The excitation features used in this study are the instantaneous fundamental frequency, the strength of excitation and the energy of excitation. The Kullback-Leibler (KL) distance is computed to measure the similarity between feature distributions of emotional and neutral speech. Based on the KL distance value between a test utterance and an utterance produced in a neutral state by the same speaker, a detection decision is made by the system. In the training of the proposed system, only three neutral utterances produced by the speaker were used, unlike in most existing emotion recognition and detection systems that call for large amounts of training data (both emotional and neutral) by several speakers. In addition, the proposed system is independent of language or lexical content. The system is evaluated using two databases of emotional speech. The performance of the proposed detection method is shown to be better than that of reference methods.

INDEX TERMS Speech analysis, paralinguistics, emotion detection, excitation source, zero frequency filtering (ZFF), linear prediction (LP) analysis, Kullback-Leibler (KL) distance.

I. INTRODUCTION

In addition to its linguistic contents, speech contains rich information about the speaker, such as the gender, age and emotional state. From these so called paralinguistic attributes, the present study addresses the last one, the speaker's emotional state, by proposing a method to automatically detect whether an utterance spoken by the speaker is emotional (e.g., produced in anger, happiness, sadness, etc.) or non-emotional. Automatic emotion detection (classifying speech as emotional vs. non-emotional) and emotion recognition (classifying the speaker's emotional state into anger, happiness, sadness etc.) helps both in human-to-human communication in speech transmission and in human-computer interaction [1]–[5]. The present study focuses on the former by studying methods to automatically detect whether utterances spoken by the speaker are emotional or non-emotional (neutral).

The associate editor coordinating the review of this manuscript and approving it for publication was Shiqing Zhang.

Inspired by their broad range of applications, automatic detection and recognition of emotions from speech has gained increasing attention in the last few years. Speech systems that are aware of speakers' emotional states can be used to implement innovations in different areas of the society [6]–[10]. For example, automatic threat detection from speech can be applied in the field of defense, detection of psychological disorders or depression from speech can be used in the assessment of mental health, and distinguishing emotional speech from neutral speech can be used to estimate customer satisfaction in call centers [6]–[8], [11]–[15].

Existing studies on the recognition/detection of emotions from speech are typically based on one of the two major approaches. First, many studies [16]–[19] have used a classical pipeline approach in which the recognition task is conducted by a system that consists of two separate parts, the front-end and the back-end. The former extracts features from speech which are used to train the classifier in the back-end to conduct the classification task. Second, a few recent studies have investigated an end-to-end approach in emotion recognition. In these systems, a deep neural network

(e.g., a convolutional neural network (CNN) or a bidirectional long short-term memory (BLSTM) network) is trained to conduct the recognition task directly from the input (either from the raw signal waveform or from the spectrogram) [20], [21]. Both of these two approaches, however, are data driven and they call for lots of training data [16]–[21]. Moreover, ideal emotion recognition/detection systems should be independent of the lexical content and language, which makes the system training even more data hungry.

Many previous studies have investigated the recognition of emotions by developing different parameterization (front-end) methods for emotional speech. These studies have shown that prosody features (pitch, intensity and duration, etc.) play a prominent role in the characterization of emotions [22]–[24]. Studies have also shown that voice quality features (jitter, shimmer, harmonic-to-noise ratio (HNR), etc.) [25], [26], spectral features (Mel frequency cepstral coefficients (MFCCs) [27], linear prediction cepstral coefficients (LPCCs), modulation spectral features (MSFs) [16], log-filter power coefficients (LFPCs) [17], line spectral frequencies (LSFs) [28], amplitude modulation cepstral coefficients (AMCCs) [29]), perceptual audio features (perceptual linear prediction coefficients (PLPs) [16]), formant frequencies [30] and their bandwidths either individually or along with prosodic features are effective in the characterization of vocal emotions [31]–[33].

The studies published in [18], [34]–[36] indicate that in machine-based emotion recognition, the system performance achieved using a few selected features may vary depending on the language and database used. Therefore, brute force approaches have been adopted in recent years by using large feature sets or subsets of features selected by feature selection algorithms [19], [37], [38]. These large feature sets typically consist of different types of features (e.g., voice quality features, spectral features, etc.) and their statistical functionals (e.g., mean, variance, min, max, range, etc.). The use of large feature sets together with their statistics might, however, suffer from the fact that emotional information may not be uniformly distributed in time, which makes the emotion detection task from continuous speech difficult.

On the other hand, studies in [24], [33], [39]–[42] investigated the emotion *detection* task instead of the emotion recognition task. In [24], [41], it was observed that robust neutral speech models were useful for discriminating different emotions. In [41], HMMs were used in an emotion detection study to generate acoustic spectral features of neutral speech. In [24], features of the pitch derived from neutral speech were used to discriminate emotions using the Kullback-Leibler distance. The gross pitch contour statistics such as mean, variance, maximum, minimum, and range were observed to be more prominent than the shape of the pitch contour. Recently, emotion detection has been studied using functional data analysis [39], [40], [42]. In this approach, energy and pitch contours of neutral speech utterances are modeled using a functional data analysis. In testing, energy and pitch contours are projected onto the reference bases, and those projections are used to discriminate emotional and neutral speech. A similar study was conducted in [43] to model the pitch contour shape of emotional speech by analyzing the falling and rising movements. One limitation of the studies in [39], [40], [42] is that all the utterances are aligned with a dynamic time warping algorithm, which is not possible in real-life applications in emotion detection.

In this study, we propose an approach to automatically detect emotional segments present in a speech signal using excitation features of speech. Instead of recognizing the emotional category of speech, the problem is formulated as emotion detection (or emotion event detection) which is close to real-life applications. The study takes advantage of three excitation features of speech (the instantaneous fundamental frequency, the strength of excitation, the energy of excitation) to measure the deviation of emotional speech from neutral speech. Using these deviations, an automatic emotion detection system was built. We also show that the proposed method is independent of the language and lexical content.

The remaining part of the article is organized as follows. Section II describes the basis for the study. Section III describes the details of the databases used in the present study. Section IV describes the feature extraction procedure and analysis of various excitation features. The developed automatic emotion detection system is presented in Section V together with the results of the study. Finally, Section VI gives a short summary of the study.

II. BASIS FOR THE PRESENT STUDY

In the production of emotional speech, the excitation characteristics of the human speech production mechanism change considerably compared to normal (neutral) speech. For example, in the production of angry speech, speakers typically use much greater sub-glottal pressure and vocal fold tension than in the production of neutral speech. Speakers are, however, not capable of keeping non-normal speech production settings sustainable for long periods of time. Therefore, the acoustical cues generated by these non-normal settings are not likely to be present in long spoken units such as sentences during their entire duration. This is particularly true for emotions of high arousal (such as anger, happiness, etc.) that typically require a large vocal effort. From the perception's point of view, the most important changes between emotional and neutral speech seem to take place at the suprasegmental level (prosody). Suprasegmental level changes are mostly learnt over a period of time and it is difficult to find consistent patterns which can form a separate set for each emotion [36], [42], [44], [45]. In the present study, the changes in the subsegmental features (derived within a glottal cycle) are examined for discriminating emotional speech from neutral speech.

III. EMOTIONAL SPEECH DATABASES

In this study, the following two databases of emotional speech are used: (1) the IIIT-H Telugu emotional speech

database [46] (to be dubbed the IIIT-H database in this article), and the Berlin emotional speech database [47] (to be dubbed the EMO-DB database in this article).

A. THE IIIT-H TELUGU EMOTIONAL SPEECH DATABASE

This database is a semi-natural database, which was recorded in Telugu (an Indian language) from students of the International Institute of Information Technology-Hyderabad (IIIT-H) [46]. The database consists of utterances produced using four emotions (anger, happiness, sadness and neutral state) by seven speakers (5 males and 2 females). All the speakers were instructed to script a text by remembering past memories and situations and to read the script aloud, which is an useful procedure to elicit emotional speech. Due to the use of this procedure, it is worth noting that the lexical contents of the recorded speech utterances are different between the speakers and emotions, as the contents depend on prior experiences of each individual speaker. The recording sessions were carried out in a laboratory environment using a microphone and electroglottography. The database consists of around 200 utterances which were recorded in 2-3 sessions for each speaker. The entire data was evaluated by 10 listeners for recognizability of the emotions. After the evaluation, a total of 130 utterances were selected consisting of 35, 27, 34 and 34 utterances in states of anger, happiness, neutrality and sadness, respectively. The average utterance duration was approximately 3 sec.

B. THE BERLIN EMOTIONAL SPEECH DATABASE

The EMO-DB database [47] was recorded in German at Technical University of Berlin. Ten professional native actors (5 males and 5 females) were asked to produce 10 sentences in seven emotions (anger, happiness, neutral state, sadness, fear, disgust and boredom). The recording was carried out in an anechoic chamber in one or more sessions. The entire data, consisting of around 800 utterances, was evaluated by 20 listeners for recognizability of the emotions. Those utterances which had a recognition rate higher than 80% and a naturalness rate higher than 60% were selected for further use. After this process, a total of 535 utterances remained consisting of 127, 71, 79, 62, 70, 51, and 81 utterances expressing anger, happiness, neutral state, sadness, fear, disgust, and boredom, respectively. The average utterance duration was approximately 3 sec.

IV. EXTRACTION OF EXCITATION FEATURES

The features related to the excitation component of the speech production mechanism were used for the development of the emotion detection system. The features used were: the instantaneous fundamental frequency (F_0), the strength of excitation (SoE) and the energy of excitation (EoE). For the extraction of these features, two signal processing methods were used. The instantaneous fundamental frequency and the strength of excitation were extracted using the zero frequency filtering (ZFF) method [48] and the energy of excitation was extracted using linear prediction analysis (LPA). All these features were computed around glottal closure instants (GCIs) obtained from the ZFF method.

A. ZERO FREQUENCY FILTERING (ZFF)

The ZFF method [48] is based on the observation that the impulse-like excitation caused by an abrupt closure of the vocal folds is reflected across all frequencies including zero frequency (0 Hz). In this method, the differentiated speech signal ($x[n] = s[n]$ - $s[n-1]$, where $s[n]$ is the speech signal) is passed through a cascade of two zero frequency resonators (i.e., a pair of poles on the positive real axis of the unit circle in the z-plane). This filtering is computed as

$$y_o[n] = \sum_{k=1}^{4} a_k y_o[n-k] + x[n], \tag{1}$$

where $a_1 = +4$, $a_2 = -6$, $a_3 = +4$, $a_4 = -1$. The resulting signal $y_o[n]$ is equivalent to integrating (or cumulatively summing) the speech signal four times in the discrete time domain. Hence, $y_o[n]$ grows/decays as a polynomial function of time. The growing/decaying trend in $y_o[n]$ is removed by subtracting the local mean (computed over the average pitch period) at each sample as follows

$$y[n] = y_o[n] - \frac{1}{2N+1} \sum_{i=-N}^{N} y_o[n+i], \tag{2}$$

where $2N+1$ corresponds to the number of samples used for the trend removal. The resulting signal ($y[n]$) is referred to as the zero frequency filtered signal. The negative-to-positive zero crossings (NPZCs) of the zero frequency filtered signal correspond to the glottal closure instants (GCIs).

B. FEATURE EXTRACTION

The zero frequency filtered signal can be considered as an approximate excitation waveform of the human speech production mechanism and therefore it can be used in the estimation of the excitation characteristics [48]. In order to quantify the excitation characteristics around the GCIs, F_0, SoE and EoE are used. By denoting GCIs in a voiced segment as $\mathcal{G} = \{g_1, g_2, \ldots, g_M\}$, where M is the number of GCIs, the computation of these three features is explained in the following separately for each feature.

1) INSTANTANEOUS FUNDAMENTAL FREQUENCY (F_0)

Pitch is one of the most important acoustic features of emotional speech [24], [49]. To capture variations in pitch, the instantaneous fundamental frequency (F_0) is computed from the locations of the GCIs. The interval between two successive GCIs corresponds to the instantaneous fundamental period or the pitch period (T_0), and its reciprocal gives the F_0. The instantaneous fundamental frequency is derived as follows

$$T_{0_{g_c}} = \frac{(g_c - g_{c-1})}{f_s}, \quad c = 2, 3, \ldots, M, \tag{3}$$

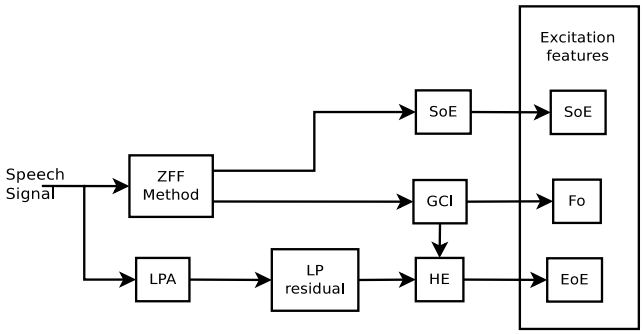

FIGURE 1. Schematic block diagram for the extraction of the excitation features.

$$F_{0_{gc}} = \frac{1}{T_{0_{gc}}} = \frac{F_s}{(g_c - g_{c-1})}, \quad c = 2, 3, \ldots, M, \quad (4)$$

where f_s is the sampling frequency.

2) STRENGTH OF EXCITATION (SoE)

SoE is defined as the slope of the zero frequency filtered signal around the c^{th} GCI and is computed as

$$SoE_{g_c} = |y[g_c + 1] - y[g_c - 1]|, \quad c = 1, 2, \ldots, M. \quad (5)$$

The SoE was shown to be useful in the analysis of vocal emotions in [46], [50], [51], where this feature was found to be large for low-arousal emotions (such as boredom and sadness), and small for high-arousal emotions (such as anger and happiness). Hence, this feature highlights the changes in relative duration of the glottal closing phase in a similar manner as the normalized amplitude quotient (NAQ) [52], [53].

3) ENERGY OF EXCITATION (EoE)

The EoE is computed from the Hilbert envelope ($HE[n]$) of the LP residual of $x[n]$ over a 1-ms region around the c^{th} GCI and is defined [50] as follows

$$EoE_{g_c} = \frac{1}{2K+1} \sum_{i=-K}^{K} HE^2[g_c + i], \quad c = 1, 2, \ldots, M, \quad (6)$$

where 2K+1 corresponds to the number of samples in a 1-ms window. This feature reflects changes in vocal effort [46], [50]. Results reported in [46], [50] indicated that the EoE is large for high-arousal emotions and small for low-arousal emotions.

The steps involved in the extraction of the selected three excitation features are shown in a schematic block diagram in Fig. 1. Illustrations of features computed from neutral and angry speech of a male speaker are shown in Figs. 2 and 3, respectively. Similarly, features extracted from neutral and angry speech of a female speaker are shown in Figs. 4 and 5, respectively. In all these figures, (a) shows the analysed speech segment, (b) shows the F_0 contour, (c) shows the SoE contour, and (d) shows the EoE contour. It can be observed both for the male and female speaker that the dynamic range

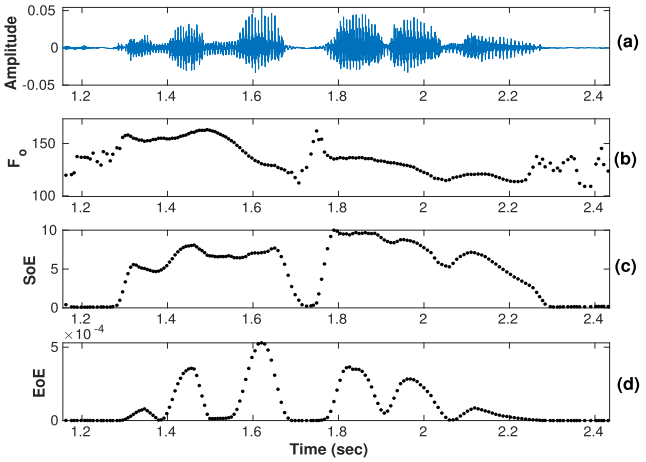

FIGURE 2. An illustration of the excitation features derived from the neutral speech of a male speaker. (a) A segment of the speech signal, (b) F_0 contour, (c) SoE contour, and (d) EoE contour.

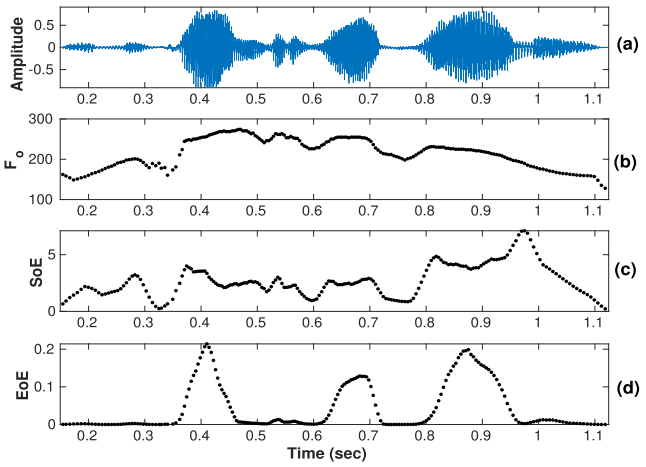

FIGURE 3. An illustration of the excitation features derived from the angry speech of a male speaker. (a) A segment of the speech signal, (b) F_0 contour, (c) SoE contour, and (d) EoE contour.

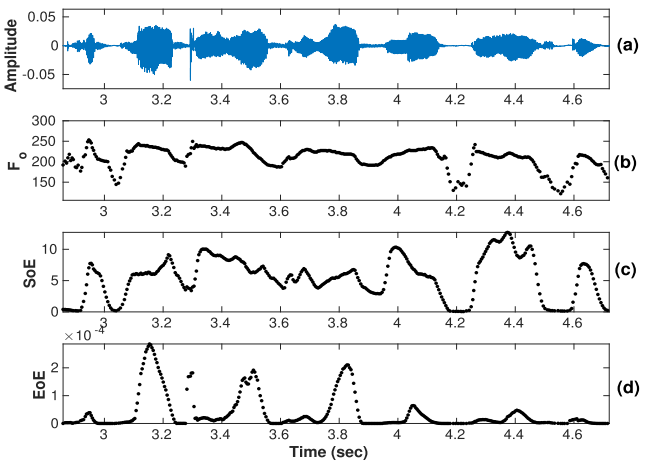

FIGURE 4. An illustration of the excitation features derived from the neutral speech of a female speaker. (a) A segment of the speech signal, (b) F_0 contour, (c) SoE contour, and (d) EoE contour.

of the feature values varies from the neutral state to anger (emotional).

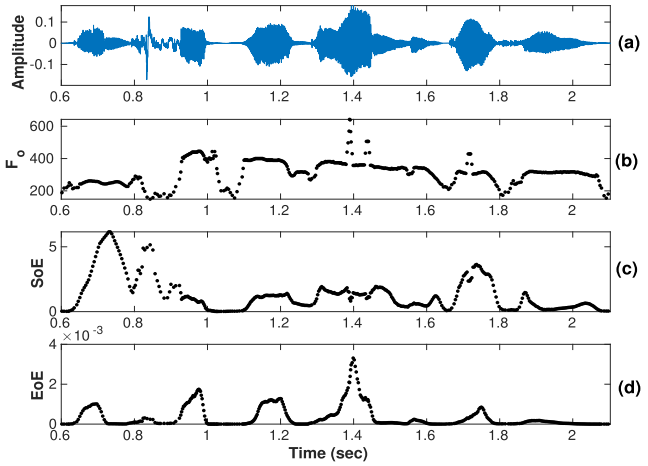

FIGURE 5. An illustration of the excitation features derived from the angry speech of a female speaker. (a) A segment of the speech signal, (b) F_0 contour, (c) SoE contour, and (d) EoE contour.

C. FEATURE ANALYSIS

From the three extracted excitation features, three 2-dimensional feature spaces are formed by combining two of the features at a time. The three feature combinations are defined as follows:

$C1 : (SoE\ vs.\ F_0)$
$C2 : (F_0\ vs.\ EoE)$
$C3 : (SoE\ vs.\ EoE).$

For an illustration, Figs. 6 and 7 show the excitation features extracted from a reference utterance (the neutral state marked by $'o'$ in red color) and from a test utterance (the angry state marked by $'*'$ in black color) using the 2-D feature spaces C1, C2 and C3 for a male and female speaker, respectively. For a better visualization, Figs. 8 and 9 show a 3-D feature space of the excitation features (with the neutral state marked by $'*'$ in red color and the angry state marked by $'*'$ in black color) for a male and female speaker, respectively. From the figures, it can be seen that the two emotional classes can be clearly distinguished from each other. The present study addresses a speaker-specific scenario, where the neutral reference utterance and emotional test utterance are from the same speaker. The representations for the reference and test utterances in C1, C2 and C3 are modeled using a Gaussian probability distribution function which is represented by mean and covariance matrices. To measure the similarity between the reference and test utterance in the 2-D feature spaces, the KL distance is computed. The KL distance is given by

$$D = \frac{1}{2}(tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^T \Sigma_1^{-1}(\mu_1 - \mu_0) - k - ln(\frac{det\,\Sigma_0}{det\,\Sigma_1})) \quad (7)$$

where, D is the KL distance, k is the dimension of the distribution, Σ_0, Σ_1 are the covariance matrices, and μ_0, μ_1 are the mean vectors of feature pair distributions of reference and test utterances, respectively.

The KL distance values are shown in Table 1 for two speakers (one male and one female) of the IIIT-H database. The table shows the KL distances between a reference utterance (*neutral state*) and a test utterance (*neutral state*, *anger*, *happiness*, and *sadness*) for each of the three 2-D feature spaces. Similarly, KL distance values are shown in Table 2 between a reference utterance (*neutral state*) and a test utterance for two speakers (one male and one female) of the EMO-DB database by analysing six emotional states (*anger*, *happiness*, *sadness*, *disgust*, *fear*, and *boredom*). From the tables, it can be observed that the KL distance values are clearly lower (i.e., distributions are more similar) when both the reference utterance and the test utterance represent the neutral state. However, the distance values are larger (i.e., the distributions are less similar) when the test utterance represents emotional speech. These observations are true for both the male and female speakers. These observations suggest that the extracted excitation features are capable of distinguishing emotional speech from speech in a neutral state. Inspired by these findings, we developed an automatic emotion detection system using the three 2-D features spaces of the excitation features.

V. EMOTION DETECTION SYSTEM AND RESULTS

The analyses reported in Section IV suggest that the absolute KL distance values can be used for the detection of emotions. It can be observed that the KL distance value between the reference neutral utterance and the test neutral utterance is very low, and that the distance value is high when the test utterance is emotional (in anger, happiness, sadness, etc.). In this study, the KL distance to be used in the detection was obtained by simply summing all the three KL distances computed using the features spaces C1, C2, and C3. This sum of the KL distance values was empirically experimented with by using two male and two female speakers from both databases. The KL distance sum value of 4.2 was chosen as a detection threshold for the IIIT-H database, i.e., a summed KL distance value smaller than 4.2 corresponded to neutral speech and a value larger than 4.2. corresponded to emotional speech. Similarly, for the EMO-DB database, the KL distance sum value of 2.8 was chosen as a detection threshold.

A block diagram of the proposed automatic emotion detection system is shown in Fig. 10. The proposed system is speaker specific, that is the system detects whether a test utterance has been spoken in an emotional state by comparing it to speech produced in a neutral state by the same speaker. Therefore, the system calls for utterances spoken in a neutral state by the test speaker for initialisation. Once the system has been initialised, the subsequent speech segments from the continuous speech of the test speaker are processed to detect the underlying emotional state (i.e., emotional vs. neutral). This is achieved by computing the summed KL distance value between the neutral speech utterances that were used in the initialisation phase and the subsequent segments of continuous speech. As the proposed approach does not call for a separate training phase with large amounts of training data,

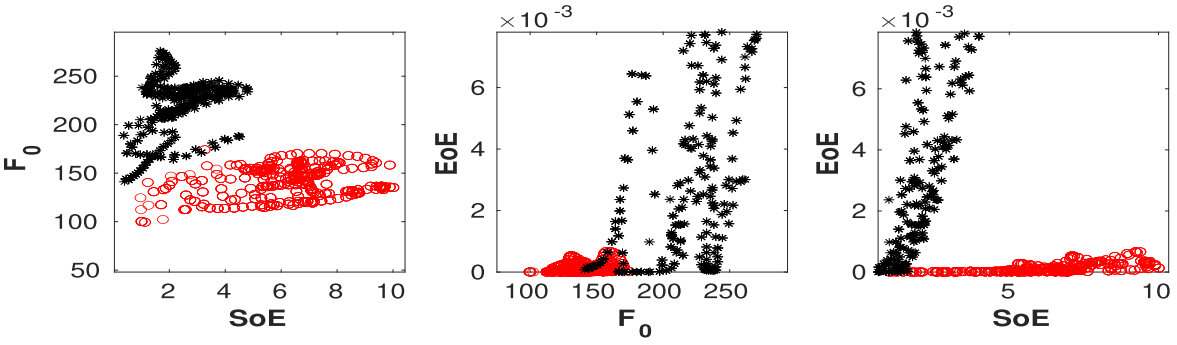

FIGURE 6. The three 2-D feature spaces for a male speaker's neutral speech (marked by 'o' in red color) and angry emotional speech (marked by '*' in black color).

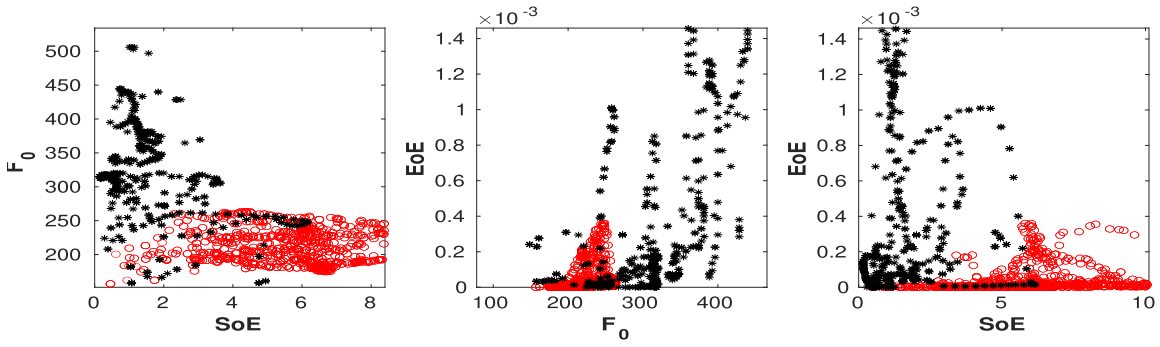

FIGURE 7. The three 2-D feature spaces for a female speaker's neutral speech (marked by 'o' in red color) and angry emotional speech (marked by '*' in black color).

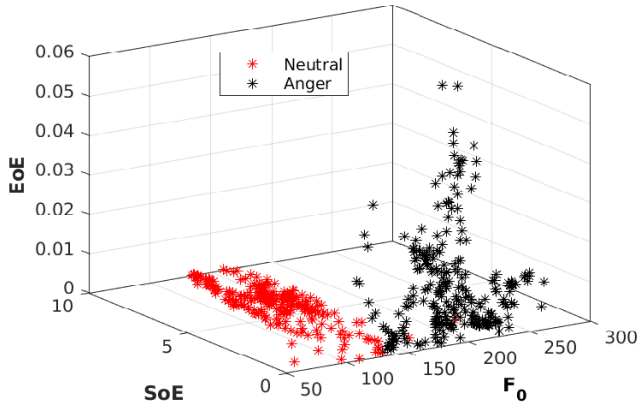

FIGURE 8. The 3-D feature space for a male speaker's neutral speech (marked by '*' in red color) and angry emotional speech (marked by '*' in black color).

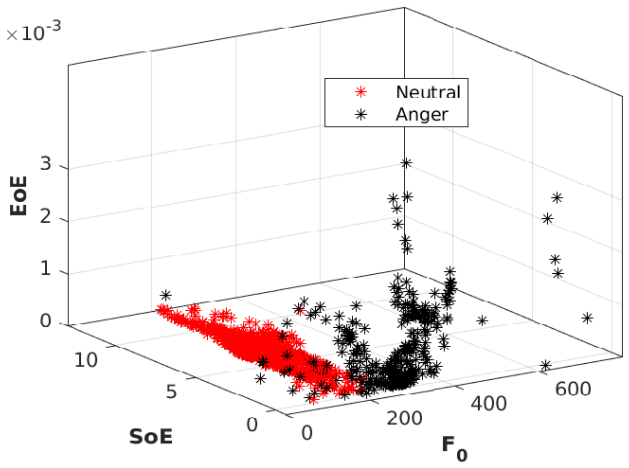

FIGURE 9. The 3-D feature space for a female speaker's neutral speech (marked by '*' in red color) and angry emotional speech (marked by '*' in black color).

and as the features used can be extracted in real-time, the method is feasible for real-life applications.

The performance of the proposed automatic emotion detection system was evaluated using the IIIT-H database and the EMO-DB database. Table 3 shows the emotion detection accuracies for the IIIT-H database. As shown in the table, the average detection accuracy is 91.67%. The accuracy in detecting a neutral state vs. anger and a neutral state vs. happiness are very high indicating that that speech in the neutral state deviates greatly from emotional speech of high arousal represented by anger and happiness. However, the accuracy in detecting the neutral state vs. sadness was lower which is explained by the reduced use of vocal effort in the production of sad speech compared to angry and happy speech. The detection accuracy computed using utterances from the EMO-DB database is presented in Table 4. In this table, the accuracies obtained by the proposed method are given in the first column and columns 2-4 present existing

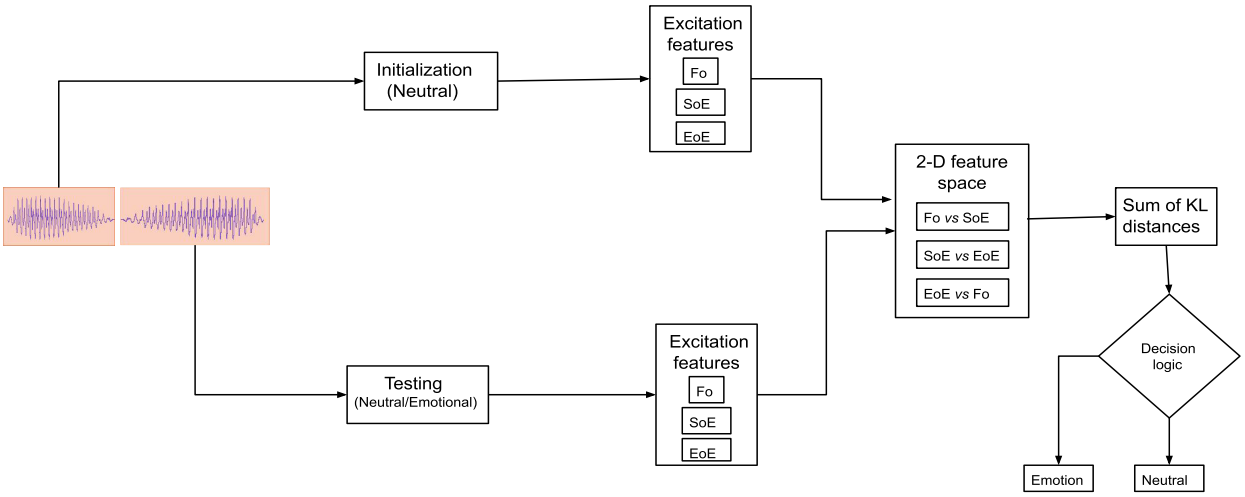

FIGURE 10. A block diagram of the proposed automatic emotion detection system.

TABLE 1. The KL distance values between the reference utterances (‘*neutral state*’) and test utterances of different emotions in the IIIT-H database.

	2-D feature space		
	C1	C2	C3
Speaker 1 (Male)			
neutral state vs. neutral state	0.643	0.434	0.157
neutral state vs. anger	9.416	454.4	634.3
neutral state vs. happiness	1.354	7.318	8.495
neutral state vs. sadness	1.481	2.237	4.445
Speaker 2 (Female)			
neutral state vs. neutral state	0.274	0.192	0.097
neutral state vs. anger	5.654	58.71	60.42
neutral state vs. happiness	2.911	207.4	203.6
neutral state vs. sadness	1.865	2.429	3.932

TABLE 2. The KL distance values between the reference utterances (‘*neutral state*’) and test utterances of different emotions in the EMO-DB database.

	2-D feature space		
	C1	C2	C3
Speaker 1 (Male)			
neutral state vs. neutral state	0.421	0.112	0.398
neutral state vs. anger	38.72	22.52	2.794
neutral state vs. happiness	26.80	17.27	1.779
neutral state vs. sadness	12.37	1.491	9.432
neutral state vs. disgust	2.214	1.091	1.624
neutral state vs. fear	8.342	6.136	0.358
neutral state vs. boredom	6.796	0.341	5.124
Speaker 2 (Female)			
neutral state vs. neutral state	0.421	0.447	0.722
neutral state vs. anger	43.24	38.78	2.664
neutral state vs. happiness	24.67	22.07	1.97
neutral state vs. sadness	8.941	0.835	9.453
neutral state vs. disgust	3.425	3.501	0.152
neutral state vs. fear	12.09	10.39	1.033
neutral state vs. boredom	30.11	0.852	30.76

TABLE 3. Emotion detection results for the IIIT-H database.

	Accuracy[%]	**Chance level**[%]
neutral state vs. anger	97.14	50
neutral state vs. happiness	92.59	50
neutral state vs. sadness	85.29	50
neutral vs. emotional	**91.67**	**50**

emotion detection results reported in [39] and [40] using the energy and pitch contour features with functional data analysis modeling. The average emotion detection accuracy using the proposed method for the EMO-DB database is 86.16%. It can also be observed from this database that the emotional states of higher activation (anger, happiness, disgust and fear) can be detected with better accuracy compared to boredom which is an emotional state of lower activation. The reduced performance of the proposed method for detecting boredom may be due to the use of lesser vocal effort in producing the utterances. This observation is in line with the studies reported in [40], [54] and [55], which showed that acoustic features in general better discriminate emotions of high activation than those of low activation. Moreover, the detection accuracies for low-activation emotions are worse, because some of the speech segments in these emotions are closer to speech in a neutral state. It is interesting to observe that the discrimination between utterances in a neutral state and those expressing sadness is high for this database. This may be due to the fact that the speech signals of the EMO-DB database were produced by actors, that is, by professional voice users who are most likely capable of expressing sad emotions using more vocal effort than, for example, the student speakers of the IIIT-H database. From Table 5, it can be observed that the proposed emotion detection system based on speech excitation features provides an improvement of approximately

TABLE 4. Emotion detection results (in [%]) between the neutral state and the six emotional classes of the EMO-DB database using the proposed method and existing methods utilizing functional data analysis (FDA) on F_0, energy and their combination [39], [40].

	Proposed	**FDA-**F_0 [39], [40]	**FDA-Energy** [39], [40]	**FDA-**F_0**+Energy**[%] [39], [40]
neutral state vs anger	98.42	77.7	95.0	92.9
neutral state vs happiness	92.95	78.9	81.8	88.6
neutral state vs sadness	91.94	66.3	68.6	70.5
neutral state vs disgust	78.43	71.1	84.2	84.0
neutral state vs fear	88.57	70.9	84.2	83.9
neutral state vs boredom	66.67	70.6	68.6	74.5

TABLE 5. Emotion detection results (in [%]) between the neutral state and emotional speech of the EMO-DB database using the baseline feature sets (F_0, energy and their combination with FDA [39], [40]) and the proposed method with the excitation features.

	FDA-F_0 [39], [40]	**FDA-Energy** [39], [40]	**FDA-**F_0**+Energy** [39], [40]	**Proposed**
Detection (%)	71.3 %	75.9 %	80.4 %	**86.16** %

4% over the recently proposed emotion detection method reported in [39] and [40], which uses both the pitch and energy contour modeling with functional data analysis (with accuracy of 80.4%) for the EMO-DB database.

The proposed emotion detection system was also tested in an on-line mode. The online emotion detection was carried out by processing 1-sec utterances. This was done by concatenating all the utterances (emotional and neutral) of the speaker for testing. The emotion detection results obtained from the online testing with the IIIT-H database and EMO-DB database are shown in Table 6. The average emotion detection accuracy for the on-line mode experiment for the IIIT-H and EMO-DB databases were 83% and 77%, respectively. The performance of the system in the on-line mode was lower compared to the evaluation with the entire utterance. This is due to some loss of suprasegmental information in a shorter speech segments. This reduction in accuracy is expected as some of the 1-sec speech segments may not contain information about the emotional state. This is also in line with the studies in [24], [49] and [39], where it was shown that the emotionally salient aspects of speech are important in emotion detection, recognition and synthesis of expressive speech.

TABLE 6. Emotion detection results obtained by processing 1-sec speech segments for the IIIT-H database and EMO-DB database.

	IIIT-H	**EMO-DB**
neutral state vs anger	88.23	86.9
neutral state vs happiness	83.10	85.2
neutral state vs sadness	77.74	78.3
neutral state vs disgust	-	71.6
neutral state vs fear	-	77.8
neutral state vs boredom	-	64.2
neutral vs. emotional	**83.02**	**77.33**

VI. SUMMARY AND CONCLUSION

In this study, we proposed an automatic emotion detection system from speech using excitation features extracted around GCIs. Using the KL distance, the system measures the deviation between the reference utterance produced in a neutral state and a test utterance of emotional speech. The results of the proposed method indicate that the excitation features seem to play a prominent role in the discrimination of emotional speech from neutral speech. Although the emotion detection system is speaker specific, it can also be made speaker independent by initialising the system with a large number of neutral voices (or using an average neutral voice) and by making the appropriate decision for the emotion detection. It is also to be noted that all the segments of the test speech utterance may not be equally important in making the decision, i.e., information describing the emotion may not be distributed uniformly in time, and hence the use of higher confidence speech segments may improve the overall emotion detection/recognition systems accuracy.

The main advantage of the proposed system is that there is no need for training the system with emotional speech. The feature combinations that are used in this study can also help in developing an emotion recognition system. Since the present study demonstrates the importance of the excitation features, it may be possible to combine features from the excitation and vocal tract system to improve the performance of automatic emotion detection/recognition systems.

REFERENCES

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[2] K. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Commun.*, vol. 40, nos. 1–2, pp. 227–256, Apr. 2003.

[3] A. Jaimes and N. Sebe, "Multimodal human-computer interaction: A survey," *Comput. Vis. Image Understand.*, vol. 108, nos. 1–2, pp. 116–134, 2007.

[4] A. D. Flockhart, E. P. Mathews, and J. Z. Taylor, "System and method for detecting emotions at different steps in a communication," U.S. Patent 8 054 964, Nov. 8, 2011.

[5] V. A. Pertrushin, "System, method and article of manufacture for detecting emotion in voice signals through analysis of a plurality of voice signal parameters," *Nov. 21*, vol. 2000, uS Patent 6,151,571.

[6] T. Athanaselis, S. Bakamidis, I. Dologlou, R. Cowie, E. Douglas-Cowie, and C. Cox, "ASR for emotional speech: Clarifying the issues and enhancing performance," *Neural Netw.*, vol. 18, no. 4, pp. 437–444, May 2005.
[7] D. Morrison, R. Wang, and L. C. De Silva, "Ensemble methods for spoken emotion recognition in call-centres," *Speech Commun.*, vol. 49, no. 2, pp. 98–112, Feb. 2007.
[8] L. Devillers, C. Vaudable, and C. Chastagnol, "Real-life emotion-related states detection in call centers: A cross-corpora study," in *Proc. INTERSPEECH*, 2010, pp. 2350–2353.
[9] P.-Y. Oudeyer, "Emotion recognition method and device," U.S. Patent 7 451 079, Nov. 11, 2008.
[10] V. A. Petrushin, "Detecting emotions using voice signal analysis," U.S. Patent 7 222 075, May 22, 2007.
[11] K. Forbes-Riley and D. Litman, "Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor," *Speech Commun.*, vol. 53, nos. 9–10, pp. 1115–1136, Nov. 2011.
[12] L. Vidrascu and L. Devillers, "Detection of real-life emotions in call centers," in *Proc. 9th Eur. Conf. Speech Commun. Technol.*, 2005, pp. 1841–1844.
[13] D. C. Dowden, R. W. Hemmeter, D. E. Herr, R. J. Piereth, S. M. Salchenberger, C. S. Sehgal, and M. K. Verma, "Automation of telephone operator assistance calls," U.S. Patent 5 163 083, Nov. 10, 1992.
[14] V. A. Petrushin, "Detecting emotion in voice signals in a call center," U.S. Patent 7 940 914, May 10, 2011.
[15] V. A. Pertrushin, "System and method for a telephonic emotion detection that provides operator feedback," U.S. Patent 6 480 826, Nov. 12, 2002.
[16] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Commun.*, vol. 53, no. 5, pp. 768–785, May 2011.
[17] I. Luengo, E. Navas, and I. Hernáez, "Feature analysis and evaluation for automatic emotion identification in speech," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 490–501, Oct. 2010.
[18] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, "Towards more reality in the recognition of emotional speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Honolulu, HI, USA, Apr. 2007, pp. 941–944.
[19] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Commun.*, vol. 53, nos. 9–10, pp. 1062–1087, Nov. 2011.
[20] Y. Li, T. Zhao, and T. Kawahara, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning," in *Proc. Interspeech*, Sep. 2019, pp. 2803–2807.
[21] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, "Direct modelling of speech emotion from raw speech," 2019, *arXiv:1904.03833*. [Online]. Available: http://arxiv.org/abs/1904.03833
[22] T. Seppänen, E. Väyrynen, and J. Toivanen, "Prosody-based classification of emotions in spoken Finnish," in *Proc. INTERSPEECH*, 2003, pp. 717–720.
[23] R. Fernandez and R. Picard, "Recognizing affect from speech prosody using hierarchical graphical models," *Speech Commun.*, vol. 53, nos. 9–10, pp. 1088–1103, Nov. 2011.
[24] C. Busso, S. Lee, and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 582–596, May 2009.
[25] I. Grichkovtsova, M. Morel, and A. Lacheret, "The role of voice quality and prosodic contour in affective speech perception," *Speech Commun.*, vol. 54, no. 3, pp. 414–429, Mar. 2012.
[26] C. Gobl, "The role of voice quality in communicating emotion, mood and attitude," *Speech Commun.*, vol. 40, nos. 1–2, pp. 189–212, Apr. 2003.
[27] T. Polzehl, A. Schmitt, F. Metze, and M. Wagner, "Anger recognition in speech using acoustic and linguistic cues," *Speech Commun.*, vol. 53, nos. 9–10, pp. 1198–1209, Nov. 2011.
[28] E. Bozkurt, E. Erzin, C. E. Erdem, and A. T. Erdem, "Use of line spectral frequencies for emotion recognition from speech," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 3708–3711.
[29] M. J. Alam, Y. Attabi, P. Dumouchel, P. Kenny, and D. D. O'Shaughnessy, "Amplitude modulation features for emotion recognition from speech," in *Proc. Interspeech*, Lyon, France, Aug. 2013, pp. 2420–2424.
[30] E. Bozkurt, E. Erzin, Ç. E. Erdem, and A. T. Erdem, "Formant position based weighted spectral features for emotion recognition," *Speech Commun.*, vol. 53, nos. 9–10, pp. 1186–1197, Nov. 2011.
[31] Y. Zhou, Y. Sun, J. Zhang, and Y. Yan, "Speech emotion recognition using both spectral and prosodic features," in *Proc. Int. Conf. Inf. Eng. Comput. Sci.*, Dec. 2009, pp. 1–4.
[32] E. Mower, M. J. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 5, pp. 1057–1070, Jul. 2011.
[33] C. Min Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 293–303, Mar. 2005.
[34] B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Trans. Affect. Comput.*, vol. 1, no. 2, pp. 119–131, Jul. 2010.
[35] C.-H. Wu and W.-B. Liang, "Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels," *IEEE Trans. Affect. Comput.*, vol. 2, no. 1, pp. 10–21, Jan. 2011.
[36] P. Gangamohan, S. R. Kadiri, and B. Yegnanarayana, "Analysis of emotional speech—A review," in *Toward Robotic Socially Believable Behaving Systems: Modeling Emotions*, vol. 1. Cham, Switzerland: Springer, 2016, pp. 205–238.
[37] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.
[38] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The INTERSPEECH 2012 speaker trait challenge," in *Proc. INTERSPEECH*, 2012, pp. 254–257.
[39] J. P. Arias, C. Busso, and N. B. Yoma, "Shape-based modeling of the fundamental frequency contour for emotion detection in speech," *Comput. Speech Lang.*, vol. 28, no. 1, pp. 278–294, Jan. 2014.
[40] J. P. Aris, C. Busso, and N. B. Yoma, "Energy and F0 contour modeling with functional data analysis for emotional speech detection," in *Proc. Interspeech*, Lyon, France, Aug. 2013, pp. 2871–2875.
[41] C. Busso, S. Lee, and S. Narayanan, "Using neutral speech models for emotional speech analysis," in *Proc. Interspeech Eurospeech*, Antwerp, Belgium, Aug. 2007, pp. 2225–2228.
[42] S. R. Kadiri, P. Gangamohan, and B. Yegnanarayana, "Discriminating neutral and emotional speech using neural networks," in *Proc. Int. Conf. Natural Lang. Process. (ICON)*, 2014, pp. 214–221.
[43] P. Astrid and W. F. Sendlmeier, "Prosodic characteristics of emotional speech: Measurements of fundamental frequency movements," in *ISCA Tutorial Res. Workshop (ITRW) Speech Emotion*, 2010, pp. 75–80.
[44] M. Airas and P. Alku, "Emotions in vowel segments of continuous speech: Analysis of the glottal flow using the normalized amplitude quotient," *Phonetica*, vol. 63, no. 1, pp. 26–46, Mar. 2006.
[45] T. Waaramaa, A.-M. Laukkanen, M. Airas, and P. Alku, "Perception of emotional valences and activity levels from vowel segments of continuous speech," *J. Voice*, vol. 24, no. 1, pp. 30–38, Jan. 2010.
[46] P. Gangamohan, S. R. Kadiri, and B. Yegnanarayana, "Analysis of emotional speech at subsegmental level," in *Proc. INTERSPEECH*, Lyon, France, Aug. 2013, pp. 1916–1920.
[47] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. INTERSPEECH*, 2005, pp. 1517–1520.
[48] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1602–1613, Nov. 2008.
[49] M. Bulut and S. Narayanan, "On the robustness of overall F0-only modifications to the perception of emotions in speech," *J. Acoust. Soc. Amer.*, vol. 123, no. 6, pp. 4547–4558, Jun. 2008.
[50] S. R. Kadiri, P. Gangamohan, S. V. Gangashetty, and B. Yegnanarayana, "Analysis of excitation source features of speech for emotion recognition," in *Proc. INTERSPEECH*, 2015, pp. 1324–1328.
[51] S. R. Kadiri, "Analysis of excitation information in expressive speech," Ph.D. dissertation, Speech Process. Lab., IIIT, Hyderabad, India, Dec. 2018.
[52] M. Airas and P. Alku, "Comparison of multiple voice source parameters in different phonation types," in *Proc. INTERSPEECH*, 2007, pp. 1410–1413.
[53] P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow," *J. Acoust. Soc. Amer.*, vol. 112, no. 2, pp. 701–710, Aug. 2002.

[54] J. H. Jeon, R. Xia, and Y. Liu, "Sentence level emotion recognition based on decisions from subsentence segments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 4940–4943.
[55] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Commun.*, vol. 53, nos. 9–10, pp. 1162–1171, Nov. 2011.

SUDARSANA REDDY KADIRI (Member, IEEE) received the B.Tech. degree in electronics and communication engineering (ECE) from Jawaharlal Nehru Technological University (JNTU), Hyderabad, India, in 2011. He did his M.S. degree by research from 2011 to 2014, and later converted to Ph.D. degree at the International Institute of Information Technology, Hyderabad (IIIT-H), India. He received the Ph.D. degree from the Department of ECE, IIIT-H, in 2018. He was a Teaching Assistant for several courses at IIIT-H from 2012 to 2018. He is currently a Postdoctoral Researcher with the Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland. His research interests include signal processing, speech analysis, speech synthesis, paralinguistics, affective computing, voice pathologies, machine learning, and auditory neuroscience.

PAAVO ALKU (Fellow, IEEE) received the M.Sc., Lic.Tech., and Dr.Sc.(Tech) degrees from the Helsinki University of Technology, Espoo, Finland, in 1986, 1988, and 1992, respectively. He was an Assistant Professor with the Asian Institute of Technology, Bangkok, Thailand, in 1993, and an Assistant Professor and a Professor with the University of Turku, Finland, from 1994 to 1999. He is currently a Professor of speech communication technology with Aalto University, Espoo. He has published more than 200 peer-reviewed journal articles and more than 200 peer-reviewed conference papers. His research interests include the analysis and parameterization of speech production, statistical parametric speech synthesis, spectral modeling of speech, speech-based biomarking of human health, and cerebral processing of speech. He serves as an Associate Editor of the *Journal of the Acoustical Society of America*.