Computers, Materials & Continua
DOI:10.32604/cmc.2021.018406
Article

Tech Science Press

Mental Illness Disorder Diagnosis Using Emotion Variation Detection from Continuous English Speech

S. Lalitha[1], Deepa Gupta[2,*], Mohammed Zakariah[3] and Yousef Ajami Alotaibi[3]

[1]Department of Electronics & Communication Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Bengaluru, India
[2]Department of Computer Science & Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Bengaluru, India
[3]Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Saudi Arabia
[*]Corresponding Author: Deepa Gupta. Email: g_deepa@blr.amrita.edu
Received: 07 March 2021; Accepted: 09 April 2021

Abstract: Automatic recognition of human emotions in a continuous dialog model remains challenging where a speaker's utterance includes several sentences that may not always carry a single emotion. Limited work with standalone speech emotion recognition (SER) systems proposed for continuous speech only has been reported. In the recent decade, various effective SER systems have been proposed for discrete speech, i.e., short speech phrases. It would be more helpful if these systems could also recognize emotions from continuous speech. However, if these systems are applied directly to test emotions from continuous speech, emotion recognition performance would not be similar to that achieved for discrete speech due to the mismatch between training data (from training speech) and testing data (from continuous speech). The problem may possibly be resolved if an existing SER system for discrete speech is enhanced. Thus, in this work the author's existing effective SER system for multilingual and mixed-lingual discrete speech is enhanced by enriching the cepstral speech feature set with bi-spectral speech features and a unique functional set of Mel frequency cepstral coefficient features derived from a sine filter bank. Data augmentation is applied to combat skewness of the SER system toward certain emotions. Classification using random forest is performed. This enhanced SER system is used to predict emotions from continuous speech with a uniform segmentation method. Due to data scarcity, several audio samples of discrete speech from the SAVEE database that has recordings in a universal language, i.e., English, are concatenated resulting in multi-emotional speech samples. Anger, fear, sad, and neutral emotions, which are vital during the initial investigation of mentally disordered individuals, are selected to build six categories of multi-emotional samples. Experimental

results demonstrate the suitability of the proposed method for recognizing emotions from continuous speech as well as from discrete speech.

Keywords: Continuous speech; cepstral; bi-spectral; multi-emotional; discrete; emotion; filter bank; mental illness

1 Introduction

A mental disorder, also called mental illness or psychiatric disorder [1], is a mental or behavioral pattern that causes significant impairment or distress in terms of personal functioning [2]. Mental disorders affect emotion, behavioral control, and cognition, and cause substantial interference in the learning ability of children as well as the functioning capability of adults at work and with their families. Mental disorders tend to originate at an early age and if not diagnosed and treated the individual suffers in a chronic recurrent manner [3]. The recent decade has witnessed a significant increase in the number of people suffering from mental illness [4–7]. Further, the COVID-19 pandemic has had an adverse effect on the mental health of people directly affected by the corona virus but also on their family members and friends as well as the general public [8–10]. Thus, there exists an urgency to advance human mental health globally, resulting in a great demand for health care professionals for diagnosis and treatment [11–13].

Patients suffering from different mental disorders typically experience certain specific emotions. Anxiety and fear are associated with individuals undergoing stress [14] and seasonal affective disorder [15]. Major depressive disordered [16] and mood disordered individuals [17] are prone to sadness and, in some cases, such individuals are always emotionally neutral and do not respond to situations that would typically cause an emotional response. Anger and fear are usually experienced by COVID-19 affected patients [18]. Borderline Personality Disorder (BPD) is a prevalent mental disorder that has an identifiable emotional component. It is reported that approximately 1.6% of the general population and 20% of the psychiatric population suffer from BPD [19]. Typically, BPD patients have rapid mood swings, tend to be emotionally unstable, and experience intense negative emotions (also referred to as affective dysregulation). People suffering with BPD do not feel the same emotion at all the times [20–23]. Apart from mental illness, individuals with medical issues, for example hormonal and heart related issues, experience fear, anger, or sad emotions [24–27]. Thus, anger, fear, sad, and neutral emotions are indicators of mental disorders and other medical conditions. If these emotions could be predicted, then this would greatly help mental healthcare professionals during an initial investigation to diagnose the ailment.

Human emotions can be detected through speech, facial expressions, gestures, electroencephalography signals and autonomic nervous system signals. Amongst these modalities, recognition of emotions using speech is more popular for data collection and speech sample processing is more convenient. During the primary investigation of a mental illness, doctors spend time counseling patients [28]. In the process of continuous conversation, the sequence of emotions experienced by the patients are vital to understand the symptoms and the associated disorder. This situation would benefit from a speech-based automated system that can continuously detect the sequence of a patient's emotions during counseling. Such a system would help doctors identify the mental illness.

Speech-based automated systems have been developed for health care [29–31]. These systems are equipped with emotional intelligence, causing mental health services to be further strengthened. Various automated systems that recognize emotions using text or multimodal analysis, i.e., a combination of text, images, and the linguistics of speech, have been designed [32–34].

However, most existing automated speech emotion recognition (SER) systems are monolingual and can recognize emotions only from discrete speech. If these systems can be further enhanced to recognize emotions from continuous speech they could be more beneficial for doctors to diagnose patients with mental illness. Such a continuous SER system is proposed in this research work.

This remainder of this paper is organized as follows. Section 2 briefly review state of the art SER systems. Section 3 outlines the proposed approach and performance measures. Experiments are described and the results are discussed in Section 4. Conclusions and suggestions for future work are presented in Section 5.

2 State-of-the-Art Models

A typical SER system processes and classifies various speech signals to recognize the embedded emotions. There exist several approaches to model emotions; however, categorical and dimensional models are most common [35–38]. Categorical models deal with discrete human emotions experienced most commonly in day-to-day life. For example, Ekman proposed six basic human emotions, i.e., anger, disgust, fear, surprise, happiness, and sadness [39]. A dimensional model interprets discrete emotion in terms of valence and arousal dimension [40]. In the literature, SER based on dimensional models is referred to as continuous emotion recognition [41–43]. In both categorical and dimensional SER models, emotion is recognized from a short duration phrase (2–4 s) for monolingual, multilingual, cross-lingual, and mixed-lingual contexts [44,45].

However, with conversational/continuous speech, speech data lasts for a longer duration, and the same emotion might not exist throughout the spoken utterance. Therefore, to deal with such situations, an SER system for continuous speech is essential. Few studies have investigated SER systems for continuous speech, and emotion databases with continuous speech are not available. Yeh et al. [46] investigated a continuous SER system using a segmentation-based approach to recognize emotions on continuous Mandarin emotion speech. Their study involved discrete emotion samples with categories of angry, happy, neutral, sad, and boredom. Multi-emotional samples with variable lengths were created by combining any two discrete emotion samples belonging to different categories, such as angry–happy, neutral–sad, and boredom–happy, resulting in a total of 10 categories. Frame-based and voiced segmentation techniques were designed to evaluate the two emotions in each voice sample multi-emotional sample. A 128-feature set included jitter, shimmer, formants, linear predictive coefficients, linear prediction cepstral coefficients, Mel Frequency Cepstral Coefficients (MFCC) and MFCC derivatives, log frequency power coefficients, Perceptual Linear Prediction (PLP), and Rasta-PLP served as the speech features. Relevant features were extracted using sequential forward and sequential backward selection methods. A weighted discrete k-nearest neighbor classifier was considered that was trained using variable length utterances created from the database [46]. Fan et al. [47] investigated a multi-scaled time window for continuous SER. Their work involved recognizing two emotions from two classes of voice samples, i.e., angry–neutral or happy–neutral samples from the Emo-dB database and a Chinese database [47]. Various MFCC features, modulation spectral features, and global statistical features were employed in experiments. The LIBSVM library was applied for classification. The training data was combined and segmented uniformly to train the classifier. System performance was compared with a baseline Hidden Markov Model (HMM) system [48]. The best results were obtained using global statistical features.

Summary and Limitation of State of the Art Approaches:

From the survey conducted it is evident that various SER systems for monolingual, multilingual, cross-lingual and mixed-lingual discrete speech have been proposed in the past decade. However, few studies have considered continuous SER. In addition, the discrete speech SER systems used segmented continuous speech to train the classifier. Dedicated segmentation methods were incorporated to detect emotion variation boundaries in continuous speech using German and Chinese language voice samples.

It would be more practical and useful if a well-established SER system that works for discrete speech could also be applied for continuous speech, To a large extent, existing discrete SER systems may not be able to capture the sequence of emotions in continuous speech due to variation in emotion boundaries of training samples (derived from discrete speech) and test samples (derived from continuous speech). To address this, if some enhancements are incorporated in the prevailing SER systems for discrete speech, then continuous emotions could be better detected. Further, the SER systems should be robust for detecting emotions from a universal language, such as English, so that it can be versatile across the globe. Such an SER system is proposed in this article.

The primary contributions of this study are as follows.

a. Unique sine filter bank-based Mel-coefficient functionals are explored to recognize speech emotion.

b. A distinctive compact cepstral and bi-spectral feature combination is proposed for effective SER.

c. The proposed SER system efficiently recognizes emotions in continuous speech as well as discrete speech using a simple uniform segmentation technique.

3 Proposed Approach

The workflow of the implemented methodology is shown in Fig. 1. The principal constituent modules include database preparation, preprocessing, speech feature extraction, classification, and post-processing.

3.1 Database Preparation

Globally, the majority of people communicate in English. Hence, the SAVEE database, which contains recordings of utterances from four male native British English speakers was selected. The focus of this work is toward recognition of emotions from continuous speech of mentally disordered individuals during counseling. Thus, angry, neutral, sad, and fear emotions are considered. In the database used, the recordings comprised fifteen phonetically balanced sentences per emotion from the standard TIMIT corpus, with an additional 30 sentences for neutral emotion [49].

Creation of Multi-Emotional Voice Samples

Here, the focus is on continuous emotion detection. Due to the lack of available continuous speech emotion samples, a database needed to be created from available discrete emotion samples. In the database under consideration each sample includes a discrete emotion of 2–4 s. In a practical situation, human emotions exist for a certain period. Thus, 3–4 samples of the same emotion class are concatenated to form a voice sample of a single emotion category, as shown in

Figs. 2 and 3. Two such voice samples from different emotion categories are concatenated to create a continuous multi-emotional speech sample, as depicted in Fig. 4. Thus, in this work, continuous speech samples are multi-emotional with a duration of 7–12 s. Five different categories of multi-emotional voice samples, i.e., angry–neutral, sad–angry, angry–fear, sad–neutral, and fear–neutral are created using Audacity [50], which is an open-source audio editor and recording application software. Any two emotions from angry, neutral, sad, and fear are considered in multi-emotional speech creation as identification of these emotions are significant in any clinical investigation of an individual thought to be suffering from a mental disorder.

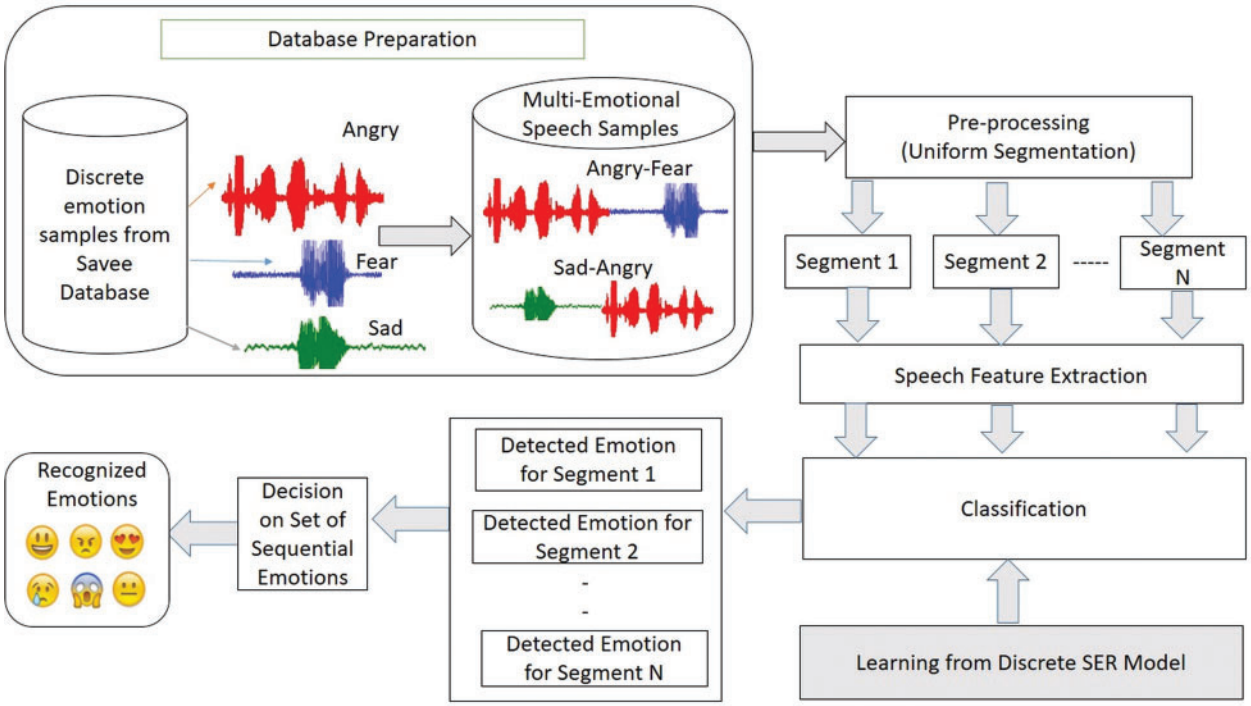

Figure 1: Proposed continuous SER work overflow

3.2 Preprocessing

This phase involves segmentation of continuous speech. As shown in Fig. 5, the speech signal is segmented uniformly into segments of constant lengths (e.g., 2 s) and two consecutive frames make an independent speech sample. Framing is performed without overlapping. Then, the emotion of each segment can be recognized.

3.3 Speech Feature Extraction

In this study, the speech feature set includes cepstral features, bispectral features, and modified sine-based MFCC coefficients.

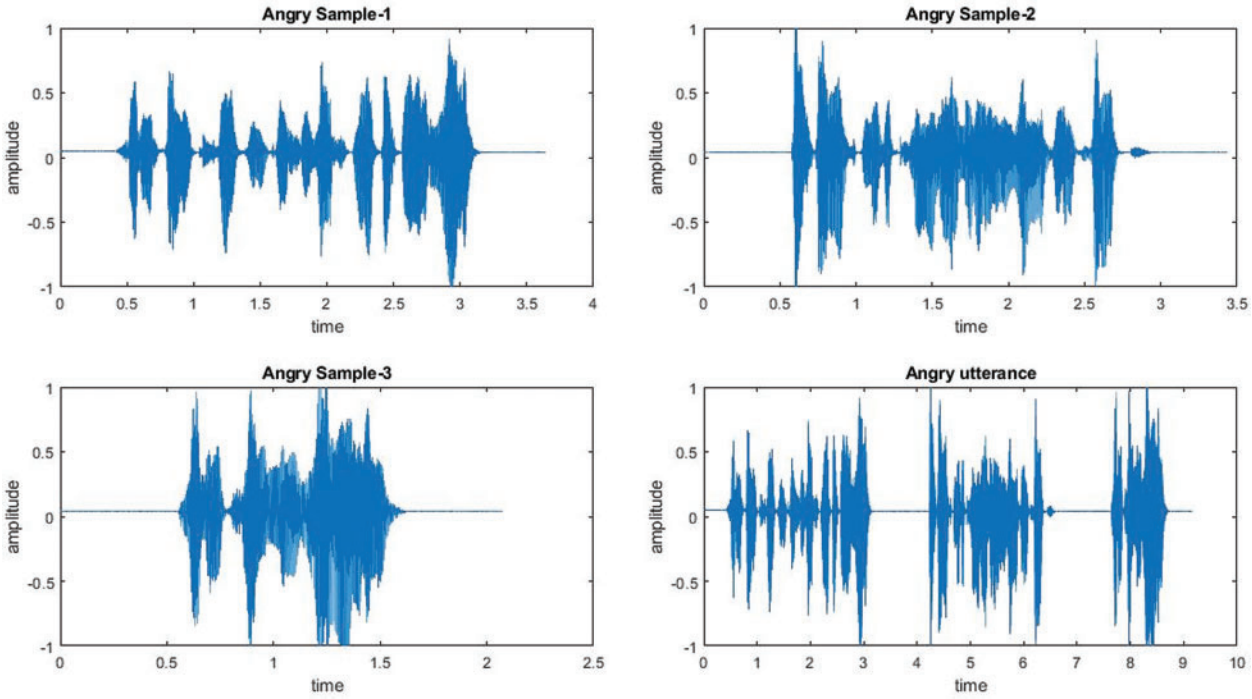

Figure 2: Creation of an angry utterance

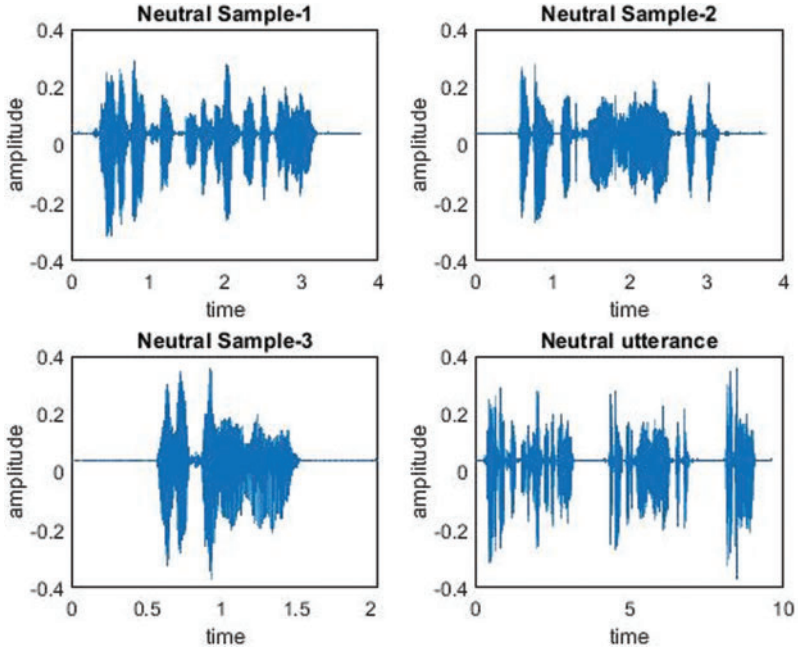

Figure 3: Creation of a neutral utterance

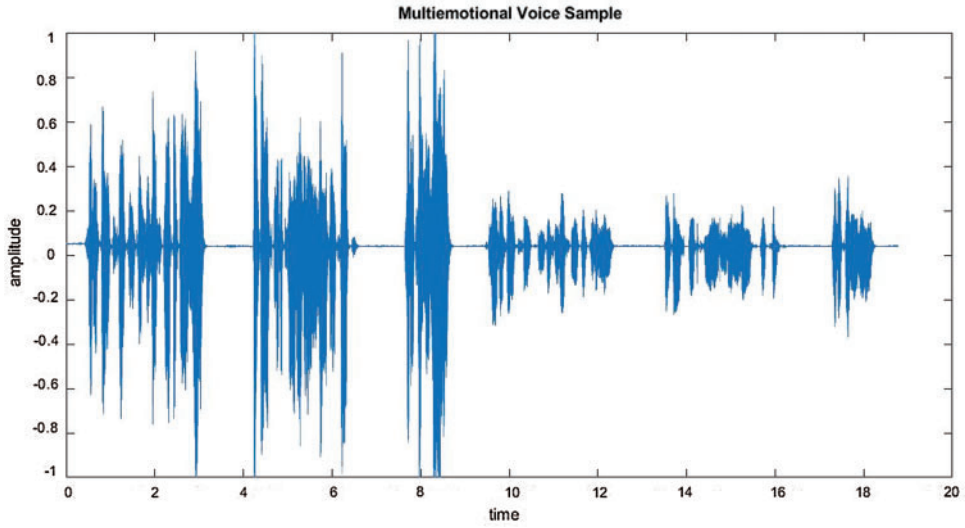

Figure 4: Combining utterances

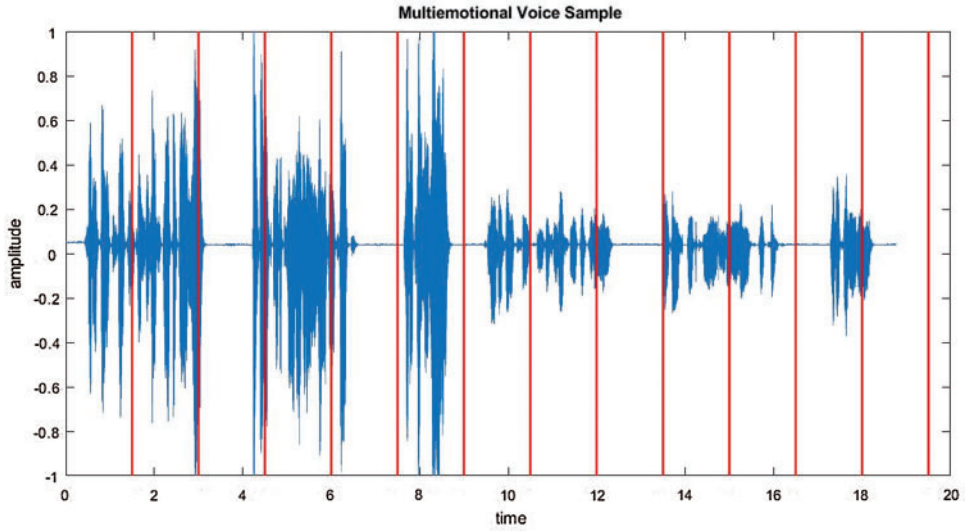

Figure 5: Uniform segmentation of continuous speech

3.3.1 Cepstral Features

Unique cepstral speech feature functionals derived from Mel, Bark, and inverted Mel filter banks along with modified H-coefficients and additional parameters are found to be quite robust for multilingual and mixed-lingual SER for discrete samples from Indian and western language backgrounds [51]. The feature set in a previous study [51] form a size of 151 coefficients, as shown in Tab. 1, which are part of the speech feature set in this work.

Table 1: Cepstral feature set [51]

Sl. No.	Speech feature	Feature size	Sl. No.	Speech feature	Feature size
1	MFCC functionals	6	11	Extended IMFCC un-voiced functionals	6
2	MFCC voiced functionals	6	12	LPC functionals	6
3	MFCC unvoiced functionals	6	13	Functionals of MEDC, LFPC, LPCC	18
4	Extended MFCC functionals	6	14	PLPC functionals	6
5	Extended MFCC voiced functionals	6	15	MFPLPC functionals	6
6	Extended MFCC un-voiced functionals	6	16	BFCC functionals	6
7	IMFCC functionals	6	17	RPCC functionals	6
8	IMFCC voiced functionals	6	18	H-Coefficients	8
9	IMFCC un-voided functionals	6	19	Functionals of cceps and rceps, rceps_ph	18
10	Extended IMFCC voiced functionals	6	20	Skewness, kurtosis, variance, frequency, phase, average amplitude, max amplitude, maximum at pitch, pitch, entropy	11

3.3.2 Bispectral Features

Fundamentally, a bispectrum is a Fourier transform of dimension two from the cumulant function of the third order, as shown in Eq. (1).

$$P(fx, fy) = E[X(fx)X(fy)X*(fx+fy)]. \tag{1}$$

Here, $P(f_x, f_y)$ denotes a bispectrum with frequencies (f_x, f_y). X(f) represents Fourier transform, * signifies complex conjugate, and E[.] means expectation of operation [52]. The bispectrum of a speech signal includes redundant data. Thus, bispectral features are selected from the non-redundant area (Ω), as shown in Fig. 6.

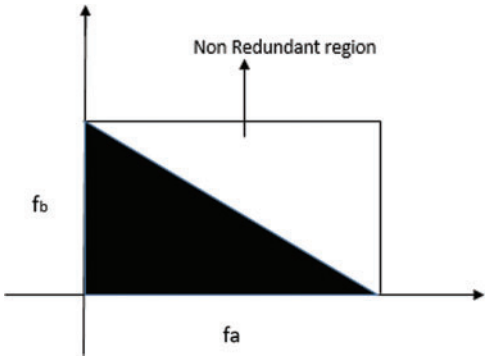

Figure 6: Non-redundant area

Frequencies represented in Fig. 6 are normalized by Nyquist frequency. Eqs. (2)–(11) illustrate the procedural steps to derive bispectral speech features. The mean magnitude of the bispectrum is expressed as follows:

$$\text{Mean Amp} = (1/\text{p}) * \sum_{\Omega} |P(fx, fy)| \tag{2}$$

where p denotes the number of points prevailing in that region [53]. The weighted center of bispectrum (WCOB) is derived using Eqs. (5)–(8).

$$g_{1d} = \frac{\sum_{\Omega} c * P(c, d)}{\sum_{\Omega} P(c, d)} \tag{3}$$

$$g_{2d} = \frac{\sum_{\Omega} d * P(c, d)}{\sum_{\Omega} P(c, d)} \tag{4}$$

$$g_{3d} = \frac{\sum_{\Omega} c * |P(c, d)|}{\sum_{\Omega} |P(c, d)|} \tag{5}$$

$$g_{4d} = \frac{\sum_{\Omega} d * |P(c, d)|}{\sum_{\Omega} |P(c, d)|}. \tag{6}$$

Here, c and d provide the bin index of the frequency existing in the region, where g_{1d}, g_{2d} represents WCOB and g_{3d}, g_{4d} are WCOB absolute values [54].

The log amplitude summation (T_a) of the bispectrum is derived as follows.

$$\text{T}_\text{a} = \sum_{\Omega} \log(|P(fx, fy|)\,. \tag{7}$$

Similarly, the log amplitude summation from diagonal elements (T_b) in the bispectrum derived as follows

$$T_b = \sum_{\Omega} log(|P(fd, fd)|)\,. \tag{8}$$

The amplitude of diagonal elements (T_c, T_d, T_e) with first and second order spectral moments is derived by:

$$T_c = \sum_{d=1}^{N} d * log(|P(fd, fd)|) \tag{9}$$

$$T_d = \sum_{d=1}^{N} (d - Tc)^2 * log\,(|P(fd, fd)|) \tag{10}$$

$$T_e = \sum_{\Omega} \sqrt{c^2 + d^2} * |P(c, d)|\,. \tag{11}$$

A total of six features comprising the bispectrum mean amplitude and five features of bispectrum log amplitudes are derived and form the part of the proposed speech feature set.

3.3.3 *Modified Sine-Based MFCC Coefficients*

The process flow for extraction of sine-based Mel coefficients is shown in Fig. 7. Initially, the power spectra of the preprocessed speech signal is derived. Differing from the conventional triangular shaped filter bank for MFCC feature extraction as discussed in an earlier SER study [55], here sinusoidal filter banks, as shown in Fig. 8, are applied to the power spectra. The center frequencies of the filter banks are given as Eq. (12).

$$\mathrm{f(p)} = \frac{N}{F_s} B^{-1}\left(p\frac{B\left(\frac{F_s}{2}\right)}{F+1}\right); \quad 1 \le p \le F \tag{12}$$

where $B^{-1}(m)$ is defined in Eq. (13) as follows:

$$B^{-1}(m) = 700(e^{\left(\frac{m}{2595}\right)} - 1) \tag{13}$$

Here, f(p) denotes center frequency, f_s represents sampling frequency, and N is the window length.

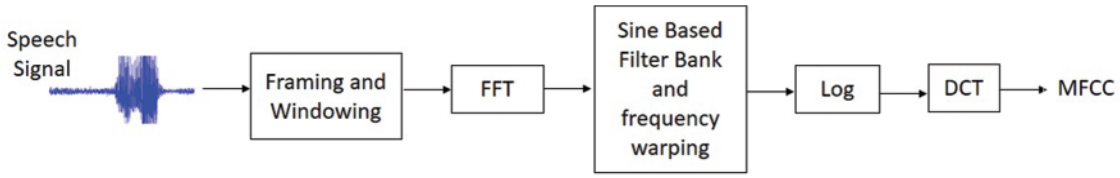

Figure 7: Sine-based MFCC feature extraction process

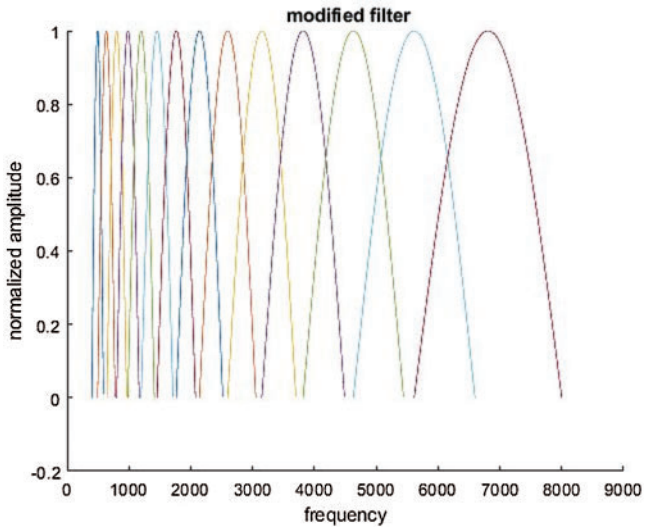

Figure 8: Sine-based filter bank

Lastly, the successive application of log and discrete cosine transform to the output of the sine-based Mel filter bank results in deriving the modified MFCC coefficients, i.e., sine-based

MFCC coefficients. Functionals of the modified MFCC, i.e., maximum, minimum, mean, standard deviation, variance, and median are considered.

In this study, for each speech signal, 151 cepstral features, six bispectral features, and six sine-based MFCC functionals (i.e., 163 coefficients) are extracted.

3.4 Classification and Post Processing

For the proposed SER work, various classifiers from Python [56] were chosen. However, compared with other classifiers, superior performance was achieved with the random forest (RF) classifier. Therefore, the RF classifier was hence considered in this work [57]. With the knowledge acquired by the classifier during training from feature vectors of discrete samples referred as learning from discrete SER Model, emotion is predicted for each continuous speech segment. The feature vector is comprised of cepstral, bi-spectral, and sine filter bank-based MFCC functionals. In the post processing phase, a decision rule is deployed to determine the sequence of emotions. For every consecutive three speech segments, the emotion predicted the maximum number of times is the emotion determined. These predicted emotions are sequences of emotions in the continuous speech.

3.5 Evaluation Metrics

In this study, performance measures of recall, precision, F-measure, and accuracy are considered to evaluate the system [58].

3.5.1 Recall

Recall *is* the number of instances that are relevant among the total number of relevant instances. Recall is also known as sensitivity.

$$\text{Recall} \quad (\%) = \frac{True\ \ Positive}{True\ \ Positive + False\ \ Negative} * 100, \tag{14}$$

where, True Positive is number of samples predicted positive that are actually positive, and False Negative is the number of examples predicted negative that are actually negative.

3.5.2 Precision

Precision gives the number of instances that are relevant among the instances retrieved. Precision is also known as the positive predictive value.

$$\text{Precision} \ (\%) = \frac{True\ \ Positive}{True\ \ Positive + False\ \ Positive} * 100. \tag{15}$$

Precision quantifies the number of correct positive predictions.

3.5.3 F-Measure

F-measure is the harmonic mean of recall and precision.

$$\text{F-measure} \ (\%) = 2 * \frac{Recall * Precision}{Recall + Precision} * 100. \tag{16}$$

3.5.4 Accuracy

Accuracy is the number of test samples of a particular emotion classified accurately with respect to the total number of test samples of the emotion under consideration.

$$\text{Accuracy } (\%) = \frac{Correctly\ recognized\ test\ samples}{Total\ number\ oftest\ samples} * 100. \tag{17}$$

4 Experimental Work, Results, and Discussion

The experimental work is performed in two successive modules. The focus of the first module involves enhancing the author's previously proposed SER system on discrete speech [51]. This is required because, although the existing SER system is suitable for recognizing emotions from various discrete speech languages, when continuous speech is tested, the performance is not similar. Therefore, the initial experimental work involves increasing the robustness of this existing SER system for discrete speech with the addition of a few more important speech features so that emotions could also be detected from continuous speech. The enhanced SER system is referred to as the proposed SER system. The second module involves experimentation on continuous speech using the proposed SER system. Both modules involve extraction of cepstral, bi-spectral and sine-based MFCC functionals speech features. An RF classifier is used. Fivefold cross-validation is applied to analyze system performance.

4.1 Module 1: Experimentation and Analysis for Proposed SER System

The previously proposed SER system for multilingual and mixed-lingual discrete speech [51] is considered in this work. The previous system comprised cepstral speech feature functionals of size 151 coefficients for each speech sample and used a simple RF classifier. Data augmentation was applied to avoid system bias toward any specific set of emotion categories. The current study focuses on recognizing emotions that are indicators of mental illness, i.e., angry, sad, fear, and neutral emotions. Thus, the initial phase of the work involved investigating the performance of the previous SER system [51] in recognizing these four emotions from discrete samples from the SAVEE database. The results obtained by this investigation are shown in Tab. 2.

Table 2: Previous SER system performance using cepstral features

Emotion	Precision (%)	Recall (%)	F-Measure (%)
Angry	90.6	96.7	93.5
Fear	77.8	70.0	73.7
Neutral	80.3	95.0	87.0
Sad	80.0	82.7	81.3
Weighted average	82.2	82.1	83.9

From Tab. 2, it can be observed that, among the four indicative emotions, the system best recognizes angry and neutral emotions with recall rates of 96.7% and 95.0%, respectively. Precision and F-Scores rates are also reported to be above 80.0% for the aforementioned emotions. The min-max rates achieved are recall 70.0%–96.7%, precision 77.8%–90.6%, and F-Score 73.7%–93.5%. In addition, weighted averages of approximately 82.0% are obtained across all performance measures. Samples from sad emotions are misclassified as neutral while fear is primarily classified as angry.

Thus, considerably lower rates are reported for sad and fear emotions. The previously proposed system has to be made more robust in recognizing fear and sad emotions along with angry and neutral emotions, such that emotions in continuous speech can be well detected.

One probable solution the authors considered to overcome this limitation was to expand the existing speech feature set and enhance system performance for fear and sad emotions. Thus, in this work, the cepstral feature set used in the previous system [51] is enhanced using bi-spectral features that capture the higher order statistics of the signal spectra. The experimental work now involves extracting cepstral–bi-spectral feature combinations and analyzing the SER system. Thus, a speech feature set of 157 coefficients (151 cepstral features and 6 bi-spectral features) for each speech sample was extracted from all the audio samples of the SAVEE database. The feature set derived from the speech samples were subjected to an emotion recognition task. The SER system performance is shown in Tab. 3.

Table 3: Performance of the discrete SER system using cepstral and bi-cepstral features

Emotion	Precision (%)	Recall (%)	F-Measure (%)
Angry	95.2	100.0	97.6
Fear	100.0	86.7	97.5
Neutral	86.4	95.0	90.5
Sad	87.2	88.3	87.7
Weighted average	91.0	92.5	93.3

From the results shown in Tab. 3, it is evident that the higher-order statistics of the bi-spectral features along with the cepstral features are significant for emotion recognition, and all emotions show performance measures greater than 85.0%. Fear and sad emotions show an increased recall rate of approximately 16% and 5%, respectively, compared with the results shown in Tab. 2. The min-max rates achieved are recall 86.7%–100.0%, precision 86.4%–100%, and F-Score 87.7%–97.6%. The min rates across the three measures, which were previously less than 80%, have improved and remained above 85%. In addition, with the inclusion of bi-spectral features, weighted averages were approximately 92.0%, which is 10% higher than those reported in Tab. 2, where only cepstral features were considered, across all performance measures. Note that, although SER performance has improved, some errors persist, i.e., sad is recognized as neutral and fear is recognized as angry. Thus, the recall rates for sad and fear emotions were between 80.0%–90.0%.

To overcome this and further enhance the emotion prediction of the SER system, the speech feature set is further expanded. For this purpose, the authors focused on altering the filter bank shape used to derive cepstral features. With an initial work in this direction, the authors considered altering one of the cepstral feature filter bank shapes proposed in Tab. 1. Among this set, MFCC has been a popular feature for various speech applications, including emotion recognition [59]. Thus, in this study, the filter bank shape of MFCC is altered. Traditionally to date, triangular filter banks have been used for MFCC feature extraction. In this work, sine-shaped filter banks have been considered, and MFCC features are derived. The extraction procedure is discussed in Section 3.

Six functionals of the Mel coefficients are derived from the sine filter bank and appended to the feature vector of the cepstral and bispectral feature combination. This resulted in a size of 163

coefficients for each speech sample. Classification was performed and the robustness of this feature combination is analyzed. The results obtained are shown in Tab. 5. With the incorporation of the new speech feature, all four emotions are optimally recognized with performance rates greater than 95% for all measures. This indicates that the shape of the filter bank has a considerable effect on the extracted Mel coefficients and hence on the emotion discriminating capability. The average accuracy of all three performance measures was 97.9%. The min-max band for recall was 95.8%–100%, precision was 96.6%–99.2%, and the F-Score was 96.2%–99.6%.

From an analysis of the results presented in Tab. 4, the previous SER system [51] is enhanced with the inclusion of bi-spectral and sine filter bank-based MFCC coefficients. This enhanced system has proven to be robust in recognizing all the four emotions of discrete speech and henceforth is referred to as the proposed SER system.

Table 4: Performance of the SER model using cepstral, bi-cepstral, and modified sine-based MFCC features

Emotion	Precision (%)	Recall (%)	F-Measure (%)
Angry	99.2	100.0	99.6
Fear	99.2	99.2	99.2
Neutral	96.6	95.8	96.2
Sad	96.7	96.7	96.7
Weighted average	97.9	97.9	97.9

4.2 Module 2: Experimentation and Analysis of Proposed SER System for Continuous Speech

In this module, experiments are conducted with regard to the recognition of emotions from continuous multi-emotional speech samples using the proposed SER system. The detailed workflow is explained in Section 3. Emotions of either angry–neutral, fear–neutral, sad–neutral, angry–fear, angry–sad, or fear–sad are included in each continuous speech sample during data creation. The feature vector of the discrete speech samples were input to train the classifier that was subsequently applied to recognize emotions in continuous speech. In this experimental procedure, for example, consider the context with recognizing emotions of an angry–neutral speech, all the samples created of this category is divided into five different folds. When continuous speech samples of a particular fold of angry–neutral is tested, all those discrete samples of angry and neutral used for data creation in that fold are removed from the discrete input to the training phase to avoid the bias during testing. The same is repeated during testing the remaining five categories of continuous speech samples.

In this context of experimentation, a multi-emotional sample of angry–neutral, as shown in Fig. 9, was tested using the proposed SER system. For each segment, the system recognizes the associated emotion. The angry emotion is denoted A, and the neutral emotion is denoted N. The decision rule was as follows: for every three consecutive segments the maximally recognized emotion is considered to be emotion. Finally, all these emotions are emotions in the continuous speech. As observed in Fig. 9, Angry-Angry-Angry–Neutral are the emotions of the speech sample tested.

All continuous emotion samples were tested, and the obtained results were analyzed. First, the performance of the proposed SER system across each fold during the fivefold cross-validation

was investigated. The bar charts in Figs. 10a–10f depict how each emotion paired with another emotion in a speech sample of each fold is recognized in continuous speech using the proposed method. Every fold consists of eight test samples across any continuous emotion category. From the plots, it is observed that both emotions in Angry–Neutral and Angry–Sad are consistently recognized from the continuous speech across all folds. However, recognition of Sad in the Sad–Neutral combination shows a large variation across the folds. Fear in the neutral or angry combination and sad combined with neutral show large variations in recognized emotions across the folds. In addition, both emotions in the Fear–Sad combination remained consistent across the folds; however, fear is confused with angry, and sad is confused with neutral, resulting in a lower recognition performance. With the investigation of emotions recognized across folds, the next step involved overall performance analysis, as illustrated in Tab. 5.

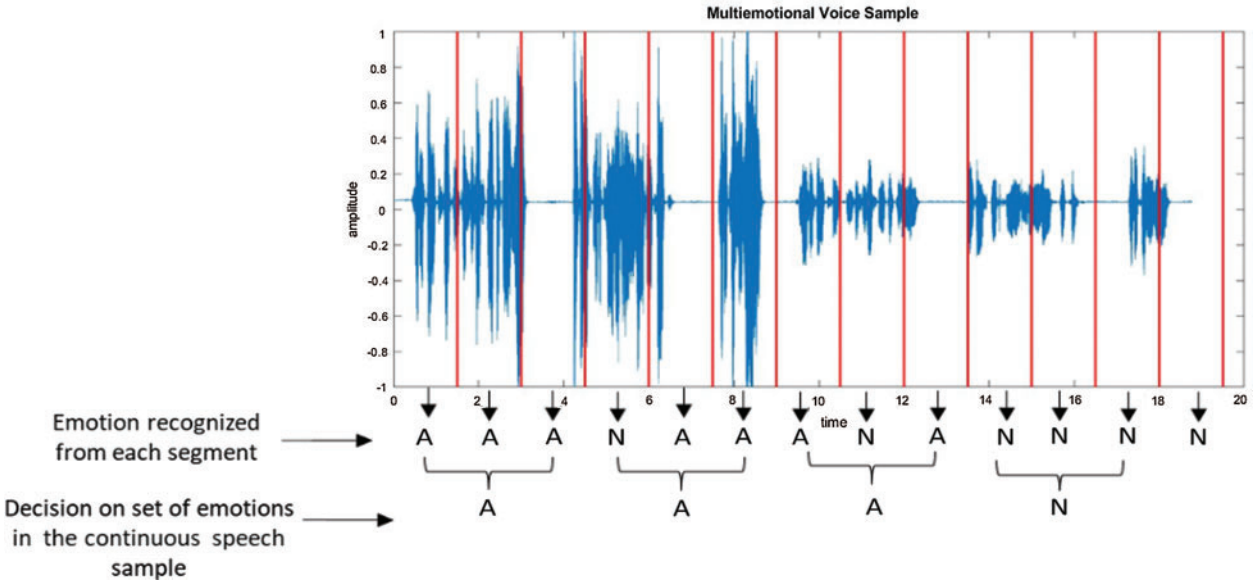

Figure 9: Recognized emotions in the continuous speech sample tested

The performance of the proposed SER system across each of these six different multi-emotional pairs in continuous speech is shown in Tab. 5. Emotions in continuous speech are better recognized from the Angry–Neutral emotion pair, with an accuracy of 85.0% and 97.0% for angry and neutral, respectively. Angry is better recognized with accuracy of at least 80.0% and higher with any of the continuous speech sample. Fear emotion is often confused with angry and is moderately recognized with Fear–Neutral and Angry–Fear scenarios. Sad is primarily classified as neutral, resulting in lower recognition performance, as observed in the Sad–Neutral combination. Thus, recognition of sad remains challenging when associated with neutral emotion. The emotions from the fear–sad continuous emotion category was found to be considerably lower, i.e., 54.6% for fear and 45.8% for sad emotion.

An analysis of the accuracy performance across the six multi-emotional categories considered in this work is shown in Fig. 11. Considerable accuracy recognition rates higher than 75.0% are guaranteed for any continuous emotion category. The Angry–Neutral emotion pair demonstrated

superior recognition rates, with an average accuracy of 91.0%. With the exception of the fear–sad combination, the min-max average accuracy band was 71.3%–91.0%.

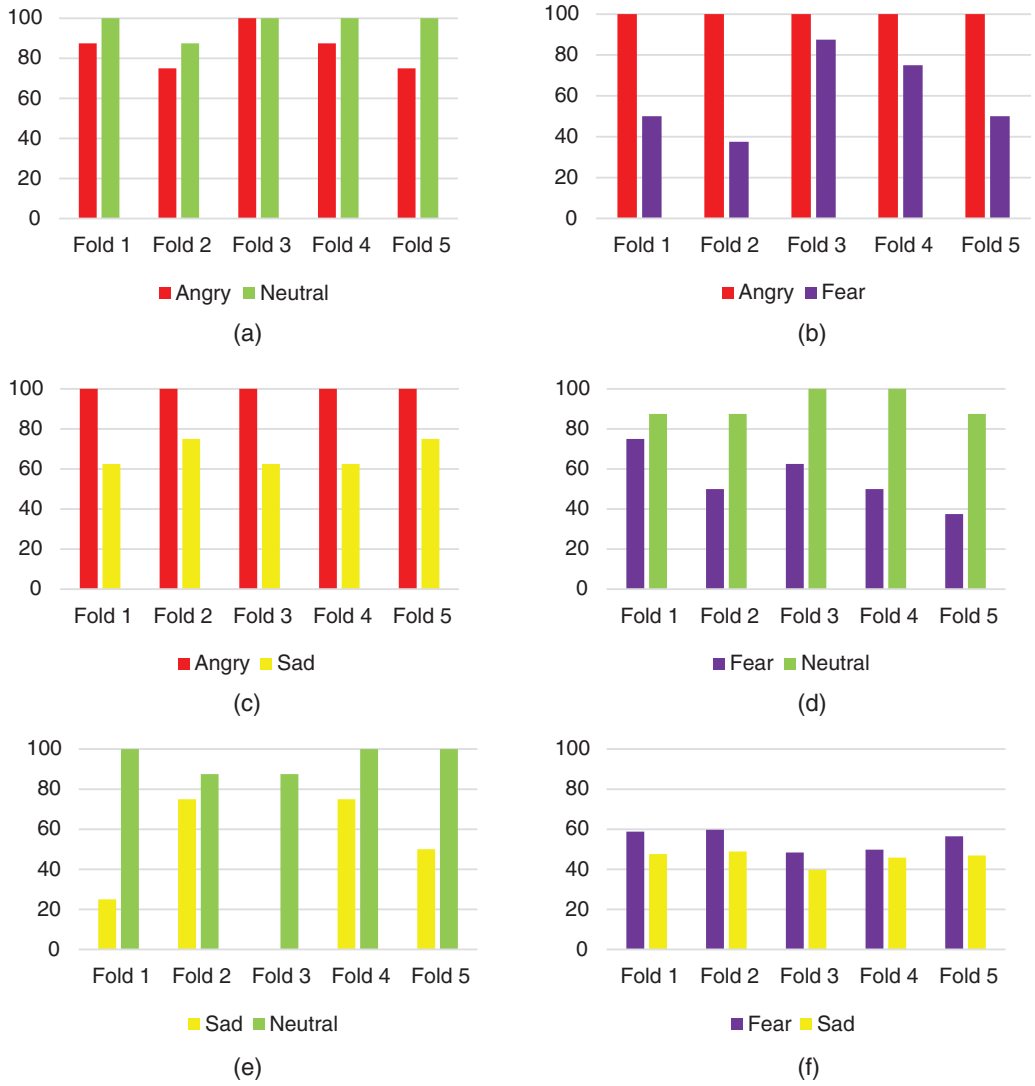

Figure 10: Accuracy (%) performance of proposed SER system for continuous speech across each fold during fivefold cross-validation (a) Angry–Neutral continuous speech (b) Angry–Fear continuous speech (c) Angry–Sad continuous speech (d) Fear–Neutral continuous speech (e) Sad–Neutral continuous speech (f) Fear–Sad continuous speech

As shown in Tab. 5, similar performance results were obtained when the first and second emotion samples of the multi-emotional combination were interchanged. For example, Angry–Neutral and Neutral–Angry continuous samples were recognized with the same accuracy by the proposed SER system.

Table 5: Accuracy (%) of the proposed SER system for continuous speech

Continuous speech emotion category	Emotion	Accuracy (%) across each emotion
Angry-Neutral	Angry	85.0
	Neutral	97.0
Fear-Neutral	Fear	61.0
	Neutral	92.5
Sad-Neutral	Sad	47.5
	Neutral	95.0
Angry-Fear	Angry	100.0
	Fear	65.0
Angry-Sad	Angry	80.0
	Sad	67.5
Fear-Sad	Fear	54.6
	Sad	45.8

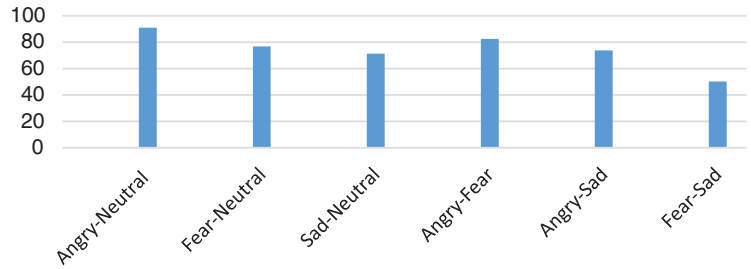

Figure 11: Average accuracy (%) of SER system for various continuous emotion combinations

4.3 Comparative Analysis of the Proposed System with Existing Continuous SER Studies

In this section, the proposed SER system is compared with existing works for recognition of emotions in continuous speech. As shown in Tab. 6, two similar studies have been identified. Though each work is validated on a different database, the comparison is performed to compare the robustness of the proposed methodology to existing techniques. The proposed work involved uniform segmentation independent of the detection of emotion variation boundaries, as performed in existing studies [46,47]. The proposed SER system showed considerable average recognition accuracy of 74.2% using a unique cepstral feature functional set of size 163. Four discrete emotions with six multi-emotional categories were involved. However, in the work of Yeh et al. [46], although five discrete emotions with 10 different multi-emotional categories using a Mandarin database are involved, with uniform segmentation, the model could achieve only 40% accuracy. Applying end point detection in the segmentation method and a feature selection method, 89.0% was achieved. Similarly, in the study conducted by Fan et al. [47], three discrete emotions with only two multi-emotional categories were involved. Though only small feature set of size 85 was applied, a multi-time scale window was applied during the segmentation stage, with an additional task of training and testing samples were chosen to be of the same length. Although both studies [46,47] demonstrated accuracy of approximately 89.0%, they only considered continuous

speech, and validation was performed on Mandarin (Chinese language voice samples) and Emo-dB (German language voice samples) databases where the speakers emotion voice recordings are highly expressive. Note that Mandarin and German are not universal languages. In contrast, the SAVEE database is considered in this work. The SAVEE database contained voice samples from male speakers in English, which is a universal language. In addition, the emotions are very flat and not expressive. The proposed SER system exhibits considerable emotion recognition for both discrete and continuous speech, proving the robustness of the emotion carrying capability of the chosen speech feature combination, which avoids detection of emotion variation boundaries, feature selection techniques, and the use of segmented continuous speech during training for continuous emotion recognition.

Table 6: Comparative analysis of the proposed continuous SER with previous studies

Author & year [Ref.]	Database/number of discrete emotions in multi-emotional voice sample/discrete emotions	Multi-emotion sample categories	Methodology [Segmentation/ features/feature vector size/ feature selection/classifier]	Average accuracy (%)
Yeh et al. 2011 [46]	Mandarin/5/Angry, neutral, sad, happy, boredom	10	Uniform segmentation or segmentation using end point detection/Cepstral and voice quality features/128/Sequential backward selection/Weighted discrete k-nearest neighbor	40.0 (uniform segmentation), 89.0 (end point detection)
Fan et al. 2014 [47]	EMO-dB/3/Angry, neutral, happy	2	Multi-time scaled window/global statistical features/85/-/Neural network	89.0
Proposed method	SAVEE/4/Angry, neutral, sad, fear	6	Uniform Segmentation/Cepstral–bispectral feature set with data augmentation/163/-/Random forest	74.2

5 Conclusion and Future Research

This study focused on the recognition of human emotions in continuous speech in a mental health context. In this study, an existing SER system for discrete speech that is quite robust for multilingual and mixed-lingual contexts is enhanced to capture emotion variations in continuous speech. It was demonstrated that altering the filter bank shape during MFCC extraction was effective in improving SER. Sine filter bank-based Mel cepstral coefficients and a cepstral–bi-spectral feature set proved to be capable of recognizing emotions from continuous speech. In addition, uniform segmentation is considered. The proposed system is independent of any dedicated segmentation techniques and feature selection algorithms. Differing from existing SER systems, the proposed system is well suited for recognizing continuous emotions in continuous speech besides discrete speech. Thus, the proposed SER system is suitable to be deployed in bots for effective mental disorder investigations.

The proposed SER system recognizes emotions from continuous English speech. Since this system is an enhanced version an existing system that is suitable for multilingual and mixed-lingual contexts, emotions from continuous speech of other languages should also be better recognized. Therefore, in future, the performance of the proposed system for continuous speech of other languages could be tested. This study is intended toward recognizing mental illness based on emotional content of speech. Therefore, real time audio recorded during counseling sessions with mental illness patients could be used to test the proposed system. In this study, two emotions are included in each multi-emotional voice sample. Future research could include more emotion categories. In addition, features could be added to the existing feature set so that the sad emotion could be better recognized in the presence of neutral or fear emotion in continuous speech. More significantly, cepstral features could be derived from different filter bank shapes.

Funding Statement: This work was partially supported by the Research Groups Program (Research Group Number RG-1439-033), under the Deanship of Scientific Research, King Saud University, Riyadh, Saudi Arabia.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] R. Oliver and T. Thayer, "Mental health disorders," *British Dental Journal*, vol. 227, no. 7, pp. 539–540, 2019.

[2] A. Risal, "Common mental disorders," *Kathmandu University Medical Journal*, vol. 9, no. 3, pp. 213–217, 2011.

[3] O. F. Norheim, "Disease control priorities third edition is published: A theory of change is needed for translating evidence to health policy," *International Journal of Health Policy and Management*, vol. 7, no. 9, pp. 771–777, 2018.

[4] S. Chancellor and M. D. Choudhury, "Methods in predictive techniques for mental health status on social media: A critical review," *NPJ Digital Medicine*, vol. 3, no. 43, pp. 1, 2020.

[5] K. Y. Huang, C. H. Wu, M. H. Su and Y. T. Kuo, "Detecting unipolar and bipolar depressive disorders from elicited speech responses using latent affective structure model," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 393–404, 2020.

[6] M. Stefanidou, C. Greenlaw and L. M. Douglass, "Mental health issues in transition-age adolescents and young adults with epilepsy," *Seminars in Pediatric Neurology*, vol. 36, pp. 100856, 2020.

[7] K. Morris and R. Edjoc, "Health fact sheet: Paraben concentrations in canadians, 2014 and 2015," *Daily (Statistics Canada)*, vol. 114, no. 12, pp. 82–625, 2018.

[8] I. I. Haider, F. Tiwana and S. M. Tahir, "Impact of the COVID-19 pandemic on adult mental health," *Pakistan Journal of Medical Sciences*, vol. 36, no. COVID19-S4, pp. 90–94, 2020.

[9] S. L. Hagerty and L. M. Williams, "The impact of COVID-19 on mental health: The interactive roles of brain biotypes and human connection," *Brain Behavior & Immunity-Health*, vol. 5, no. 9, pp. 100078, 2020.

[10] X. Ouyang, J. Huo, L. Xia, F. Shan, J. Liu *et al.,* "Dual-sampling attention network for diagnosis of COVID-19 from community acquired pneumonia," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2595–2605, 2020.

[11] A. C. Das, A. Roy and M. S. I. Salam, "Potential factors of mental health challenges during COVID-19 on the young people in Dhaka, Bangladesh," *Advanced Journal of Social Science*, vol. 7, no. 1, pp. 109–117, 2020.

[12] H. B. Turkozer and D. Ongur, "A projection for psychiatry in the post-COVID-19 era: Potential trends, challenges, and directions," *Molecular Psychiatry*, vol. 25, no. 10, pp. 2214–2219, 2020.

[13] L. Samartzis and M. A. Talias, "Assessing and improving the quality in mental health services," *International Journal of Environmental Research and Public Health*, vol. 17, no. 1, pp. 249, 2020.
[14] S. R. Bandela and T. K. Kumar, "Stressed speech emotion recognition using feature fusion of teager energy operator and MFCC," in *8th Int. Conf. on Computing, Communication and Networking Technologies*, Delhi, India, pp. 1–5, 2017.
[15] S. Melrose, "Seasonal affective disorder: An overview of assessment and treatment approaches," *Depression Research and Treatment*, vol. 2015, no. 1, pp. 1–6, 2015.
[16] N. Mahendran, P. M. D. R. Vincent, K. Srinivasan, V. Sharma and D. K. Jayakody, "Realizing a stacking generalization model to improve the prediction accuracy of major depressive disorder in adults," *IEEE Access*, vol. 8, pp. 49509–49522, 2020.
[17] F. Deligianni, Y. Guo and G. Z. Yang, "From emotions to mood disorders: A survey on gait analysis methodology," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 6, pp. 2302–2316, 2019.
[18] P. Kene, "Mental health implications of the COVID-19 pandemic in India," *Psychological Trauma: Theory, Research, Practice and Policy*, vol. 12, pp. 585–587, 2020.
[19] W. D. Ellison, L. K. Rosenstein, T. A. Morgan and M. Zimmerman, "Community and clinical epidemiology of borderline personality disorder," *Psychiatric Clinics of North America*, vol. 41, no. 4, pp. 561–573, 2018.
[20] M. Foxhall, C. H. Giachritsis and K. Button, "The link between rejection sensitivity and borderline personality disorder: A systematic review and meta-analysis," *British Journal of Clinical Psychology*, vol. 58, no. 3, pp. 289–326, 2019.
[21] E. Driessen and S. D. Hollon, "Cognitive behavioral therapy for mood disorders: Efficacy, moderators and mediators," *The Psychiatric Clinics of North America*, vol. 33, no. 3, pp. 537–555, 2010.
[22] J. Soler, D. Vega, M. Elices, A. F. Soler, A. Soto *et al.,* "Testing the reinforcement sensitivity theory in borderline personality disorder compared with major depression and healthy controls," *Personality and Individual Differences*, vol. 61, pp. 43–46, 2014.
[23] F. Kulacaoglu and S. Kose, "Borderline personality disorder (BPD): In the midst of vulnerability, chaos, and awe," *Brain Sciences*, vol. 8, no. 11, pp. 201, 2018.
[24] E. S. Paul, S. Sher, M. Tamietto and M. Mendl, "Towards a comparative science of emotion: Affect and consciousness in humans and animals," *Neuroscience and Biobehavioral Reviews*, vol. 108, no. 205, pp. 749–770, 2019.
[25] C. D. Spielberger and E. C. Reheiser, "Assessment of emotions: Anxiety, anger, depression, and curiosity," *Applied Psychology: Health and Well-Being*, vol. 1, no. 3, pp. 271–302, 2009.
[26] J. A. Arias, C. Williams, R. Raghvani, M. Aghajani, S. Baez *et al.,* "The neuroscience of sadness: A multidisciplinary synthesis and collaborative review," *Neuroscience and Biobehavioral Reviews*, vol. 111, no. Suppl 2, pp. 199–228, 2020.
[27] S. Lalitha and S. Tripathi, "Emotion detection using perceptual based speech features," in *Proc. 2016 IEEE Annual India Conf.*, Bangalore, India, pp. 1–5, 2016.
[28] I. Hassan, R. McCabe and S. Priebe, "Professional-patient communication in the treatment of mental illness: A review," *Communication and Medicine*, vol. 4, no. 2, pp. 141–152, 2007.
[29] M. Bates, "Health care chatbots are here to help," *IEEE Pulse*, vol. 10, no. 3, pp. 12–14, 2019.
[30] S. Latif, J. Qadir, A. Qayyum, M. Usama and S. Younis, "Speech technology for healthcare: Opportunities, challenges, and state of the art," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 342–356, 2021.
[31] A. N. Vaidyam, H. Wisniewski, J. D. Halamka, M. S. Kashavan and J. B. Torous, "Chatbots and conversational agents in mental health: A review of the psychiatric landscape," *Canadian Journal of Psychiatry. Revue Canadienne de Psychiatrie*, vol. 64, no. 7, pp. 456–464, 2019.
[32] S. Poria, N. Majumder, R. Mihalcea and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100943–100953, 2019.
[33] M. Bhargava, R. Varshney and R. Anita, "Emotionally intelligent chatBot for mental healthcare and suicide prevention," *International Journal of Advanced Science and Technology*, vol. 29, no. 6, pp. 2597–2605, 2020.

[34] K. Oh, D. Lee, B. Ko and H. Choi, "A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation," in *Proc. 18th IEEE Int. Conf. on Mobile Data Management*, Daejeon, pp. 371–375, 2017.

[35] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang *et al.,* "Attention-enhanced connectionist temporal classification for discrete speech emotion recognition," in *Interspeech 2019*, Graz, Austria, pp. 206–210, 2019.

[36] Z. Yao, Z. Wang, W. Liu, Y. Liu and J. Pan, "Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN," *Speech Communication*, vol. 120, no. 3, pp. 11–19, 2020.

[37] S. S. Poorna and G. J. Nair, "Multistage classification scheme to enhance speech emotion recognition," *International Journal of Speech Technology*, vol. 22, no. 2, pp. 327–340, 2019.

[38] B. T. Atmaja and M. Akagi, "Dimensional speech emotion recognition from speech features and word embeddings by using multitask learning," *APSIPA Transactions on Signal and Information Processing*, vol. 9, pp. 2825, 2020.

[39] K. Wolf, "Measuring facial expression of emotion," *Dialogues in Clinical Neuroscience*, vol. 17, no. 4, pp. 457–462, 2015.

[40] A. Yazdani, E. Skodras, N. Fakotakis and T. Ebrahimi, "Multimedia content analysis for emotional characterization of music video clips," *EURASIP Journal on Image and Video Processing*, vol. 26, pp. 1–10, 2013.

[41] Z. Huang and J. Epps, "An investigation of partition-based and phonetically-aware acoustic features for continuous emotion prediction from speech," *IEEE Transactions on Affective Computing*, vol. 11, no. 4, pp. 653–668, 2020.

[42] J. Huang, J. Tao, B. Liu, Z. Lian and M. Niu, "Multimodal transformer fusion for continuous emotion recognition," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, Barcelona, Spain, pp. 3507–3511, 2020.

[43] Z. Zhang, J. Han, E. Coutinho and B. Schuller, "Dynamic difficulty awareness training for continuous emotion prediction," *IEEE Transactions on Multimedia*, vol. 21, no. 5, pp. 1289–1301, 2019.

[44] H. Meng, T. Yan, F. Yuan and H. Wei, "Speech emotion recognition from 3D log-mel spectrograms with deep learning network," *IEEE Access*, vol. 7, pp. 125868–125881, 2019.

[45] D. Praveena and D. Govind, "Significance of incorporating excitation source parameters for improved emotion recognition from speech and electroglottographic signals," *International Journal of Speech Technology*, vol. 20, no. 4, pp. 787–797, 2017.

[46] J. H. Yeh, T. L. Pao, C. Y. Lin, Y. W. Tsai and Y. T. Chen, "Segment-based emotion recognition from continuous Mandarin Chinese speech," *Computers in Human Behavior*, vol. 27, no. 5, pp. 1545–1552, 2011.

[47] Y. Fan, M. Xu, Z. Wu and L. Cai, "Automatic emotion variation detection in continuous speech," in *Proc. Signal and Information Processing Association Annual Summit and Conf.*, Siem Reap, Cambodia, pp. 1–5, 2014.

[48] H. B. Kang, "Affective content detection using HMMs," in *Eleventh ACM Int. Conf. on Multimedia*, New York, NY, USA, pp. 259–262, 2003.

[49] H. Alshamsi, V. Kepuska, H. Alshamsi and H. Meng, "Automated speech emotion recognition on smart phones," in *9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conf.*, New York, USA, pp. 44–50, 2018.

[50] Franklin, "The sheer audacity: How to get more, in less time, from the audacity digital audio editing software," in *IEEE Int. Professional Communication Conf.*, Saragota Springs, NY, USA, pp. 92–105, 2006.

[51] S. Lalitha, D. Gupta, M. Zakariah and Y. A. Alotaibi, "Investigation of multilingual and mixed-lingual emotion recognition using enhanced cues with data augmentation," *Applied Acoustics*, vol. 170, no. 1, pp. 107519, 2020.

[52] J. Muthuswamy, D. L. Sherman and N. V. Thakor, "Higher-order spectral analysis of burst patterns in EEG," *IEEE Transactions on Bio-Medical Engineering*, vol. 46, no. 1, pp. 92–99, 1999.

[53] T. T. Ng, S. F. Chang and Q. Sun, "Blind detection of photomontage using higher order statistics," in *Int. Symp. on Circuits and Systems*, Vancouver, BC, Canada, vol. 5, pp. 688–691, 2004.

[54] X. Du, S. Dua, R. U. Acharya and C. K. Chua, "Classification of epilepsy using high-order spectra features and principle component analysis," *Journal of Medical Systems*, vol. 36, no. 3, pp. 1731–1743, 2012.

[55] S. Lalitha, S. Tripathi and D. Gupta, "Enhanced speech emotion detection using deep neural networks," *International Journal of Speech Technology*, vol. 22, no. 3, pp. 497–510, 2019.

[56] Z. Dobesova, "Programming language Python for data processing," in *Int. Conf. on Electrical and Control Engineering*, Yichang, China, pp. 4866–4869, 2011.

[57] F. Noroozi, T. Sapinski, D. Kaminska and G. Anbarjafari, "Vocal-based emotion recognition using random forests and decision tree," *International Journal of Speech Technology*, vol. 20, no. 2, pp. 239–246, 2017.

[58] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and f-ccore, with implication for evaluation," *Lecture Notes in Computer Science*, vol. 3408, pp. 345–359, 2005.

[59] S. Lalitha and D. Gupta, "An encapsulation of vital non-linear frequency features for speech applications," *Journal of Computational and Theoretical Nanoscience*, vol. 17, no. 1, pp. 303–307, 2018.