




DNN-HMM-Based Speaker-Adaptive Emotion Recognition Using MFCC and Epoch-Based Features

Md. Shah Fahad¹  · Akshay Deepak¹ · Gayadhar Pradhan² · Jainath Yadav³

Received: 23 April 2019 / Revised: 12 June 2020 / Accepted: 13 June 2020 / Published online: 21 July 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

Speech emotion recognition (SER) systems are often evaluated in a speaker-independent manner. However, the variation in the acoustic features of different speakers used during training and evaluation results in a significant drop in the accuracy during evaluation. While speaker-adaptive techniques have been used for speech recognition, to the best of our knowledge, they have not been employed for emotion recognition. Motivated by this, a speaker-adaptive DNN-HMM-based SER system is proposed in this paper. Feature space maximum likelihood linear regression technique has been used for speaker adaptation during both training and testing phases. The proposed system uses MFCC and epoch-based features. We have exploited our earlier work on robust detection of epochs from emotional speech to obtain emotion-specific epoch-based features, namely instantaneous pitch, phase, and the strength of excitation. The combined feature set improves on the MFCC features, which have been the baseline for SER systems in the literature by +5.07% and over the state-of-the-art techniques by +7.13%. While using just the MFCC features, the proposed model improves upon the state-of-the-art techniques by 2.06%. These results bring out the importance of speaker adaptation for SER systems and highlight the complementary nature of the MFCC and epoch-based features for emotion recognition using speech. All experiments were carried out on an IEMOCAP emotional dataset.

Keywords Emotion recognition · Epoch-based features · Deep neural network (DNN) · Gaussian mixture model (GMM) · Hidden Markov model (HMM) · Speaker-adaptive training (SAT) · Zero-time windowing (ZTW)

✉ Md. Shah Fahad
shah.cse16@nitp.ac.in

Extended author information available on the last page of the article

1 Introduction

The emotional state of a person can be identified using data sources such as speech, text, facial expression, brain signal (EEG), and a combination of two or more of these [1,5,19]. Here, we focus on emotion recognition using speech because it is the most natural way of communication and easy to collect. Speech emotion recognition (SER) has attracted a lot of attention in the research community due to its applicability in a multitude of real-life contexts, e.g., human–computer interaction systems such as interactive movies [24], storytelling and E-tutoring applications [34], retrieval and indexing of the video/audio files [31], improving interaction at call centers [17], assisting psychological treatments [27], and surveillance systems [6].

Speech is produced as a combination of activities in the vocal tract system and glottal source. Based on this, the features obtained from speech can be broadly sub-grouped under two categories: (i) the system features that are related to the vocal tract (e.g., mel-frequency cepstral coefficients (MFCCs), linear predictive cepstral coefficients (LPCCs), and their derivatives) and (ii) the excitation source features that are related to the glottal source (e.g., pitch, phase of the linear prediction residual signal, and strength of glottal closure instants). Apart from this, the prosodic features are also widely used. They are derived from the changes in the speech attributes with respect to time (e.g., jitter, shimmer, and duration) [37,40]. The speech attributes may be related to both the vocal tract and glottal source. Most of the existing works in SER have been based on the system [8,38] and prosodic features [21,26]. The combination of system and prosodic features has been widely used for emotion recognition. For example, Ververidis et al. [35] used suprasegmental features such as energy, F0, formant locations, energy, dynamics of F0, and formant contours for emotion classification. The statistical parameters of F0 such as the maximum, minimum, and median values and the slopes of F0 contours have emotion-specific information [7,39]. Wang et al. [38] and Nicholson et al. [26] combined the prosodic and system features for emotion classification. These results show that the features (prosodic and system features) containing complementary information can significantly improve the performance of SER systems.

Though the pitch-based features are widely used for emotion recognition, these features were not derived from the epoch locations. An epoch is the glottal closure instant at which the excitation of the vocal tract is maximum. In this work, pitch and other excitation source features have been derived from the epoch locations. Further, the epoch-based features have been combined with the MFCC features. To the best of our knowledge, Krothapalli and Koolagudi's [15] work is the only exception that combined two emotion recognition models developed using the MFCC and epoch-based features. They developed the models using auto-associative neural networks and support vector machines. The experiments were carried out on the IITKGP-SESC dataset. The accuracy of the combined model significantly increased with respect to the individual models. The zero-frequency filtering (ZFF) method was used for detecting the epoch locations. However, the accuracy of epoch detection using the ZFF method is not satisfactory for emotional speech because it requires a priori pitch period to detect the epoch locations [41]. In this work, the epoch locations are detected by the zero-time windowing (ZTW) method [41], which is robust for emotional speech.

The behavior of the vocal cords changes rapidly in an emotional speech. The epoch-based features capture this behavior of the vocal cords. The MFCC and prosodic features are computed using block processing, while the epoch-based features are computed at each epoch location. The advantage that the epoch-based features offer over the MFCC and prosodic features is that they are better at capturing the rapid variations in an emotional speech. This is because the MFCC and prosodic features capture frame-level information, whereas the epoch-based features capture information at each epoch location. This is more fine-grained because multiple epoch locations can be present within a frame.

SER systems typically perform well when the same set of speakers are used for both training and testing phases. However, in real life, most SER systems are more likely to encounter a new test speaker on which the system has not been trained. In such cases, i.e., training and testing with different speakers, the performance degrades dramatically. This is because of the anatomical and morphological variation in the vocal-tract geometry of different speakers. To deal with this challenge, two popular approaches exist in the literature: (i) speaker normalization [4,33] and (ii) speaker adaptation [10]. While a few research works have used speaker normalization for developing SER systems [4,33], to the best of our knowledge, speaker adaptation has not been used for SER. Mariooryad and Busso [20] used various acoustic features highlighting their dependencies on speakers, emotions, and lexical contents. They further normalized speakers and lexical factors for SER. On the other hand, speaker adaptation has been used for speech recognition [10], but not for SER. Speaker adaptation techniques explicitly use speaker information to develop a robust speaker-independent SER system.

The performance evaluation of SER systems in real-life scenarios requires an appropriate dataset. Most of the SER systems have been developed using acted datasets. Therefore, they fail to detect emotions in natural utterances. The emotion distributions in the acted and natural speech do not match because an acted speech is recorded in a restricted environment, thereby lacking the variations in natural speech. Some of such restricted qualities are the near-constant length of the utterances, limited text, and exaggeration of emotions. The IEMOCAP [3] dataset deals reasonably with these constraints to incorporate naturalness. Hence, there is a significant gap in the accuracy of SER systems when evaluated on the IEMOCAP dataset as compared to the other acted datasets (such as [2]). Due to its naturalness, the IEMOCAP dataset is frequently used by researchers to evaluate their works.

The most recent works on SER using the IEMOCAP dataset are [12,20,23]

In [20], various acoustic features and their dependencies across emotions, speakers, and phonemes are analyzed using factor analysis. Further, the speaker and phoneme factors are normalized using whitening transformation. The remaining approaches (i.e., [12,23]) are deep-learning-based models. These deep-learning models have been developed using either the hand-crafted features (MFCC, pitch, voice quality, etc.) or a raw spectrogram. Both of these approaches for feature extraction produce nearly the same accuracy. However, the hand-crafted features require significant manual labor and knowledge expertise. In [12], a DNN model was used to extract features from speech segments (referred to as local features). Further, utterance-level features (referred to as global features) were constructed using the statistics of the posterior

probabilities. These utterance-level features were fed to the extreme learning machine (ELM) classifier. In [23], the raw spectrogram and the low-level descriptors (LLDs) features were modeled with attentive long short-term memory (LSTM). The intuition behind attentive LSTM is that all the frames do not contribute to emotion recognition; greater attention is learned for the contributing frames using attentive LSTM.

Based on the above discussion, the following are the motivations for our work:

- **Exploring the complementary nature of the MFCC and epoch-based features** The epoch-based features (categorized under source features) in combination with the MFCC features have been less explored for developing SER systems. Further, ZTW method [41] for detecting the epoch locations has not been used before for extracting the epoch-based features. The ZTW method has been shown to be robust to the variation of emotion in speech. Another reason to explore the epoch-based features is that they are better at capturing the rapid variations in an emotional speech because they collect information at the epoch locations, which are more fine-grained than block processing.
- **Speaker adaptation** Though most of the proposed SER systems in the literature have been evaluated in a speaker-independent manner, they do not use speaker adaptation during training or testing. On the other hand, DNN-HMM-based speaker-adaptive technique is very popular in speech recognition systems [10], where feature space maximum likelihood linear regression (fMLLR) is used for speaker adaptation. However, to the best of our knowledge, speaker adaptation has not been used for developing SER systems.

The remaining part of the paper is structured as follows: Sect. 2 describes the detection of the epoch-based features and briefly discusses the MFCC features. Section 3 discusses the proposed SER system. The description of the IEMOCAP dataset is given in Sect. 4. The experimental setup and results are discussed in Sect. 5. Section 6 concludes the paper with future directions.

2 Features

In this section, we describe the extraction of the epoch-based features using the ZTW method [41] and the extraction of the MFCC features. While the MFCC features can be extracted from both the voiced and unvoiced regions, the epoch-based features are applicable only for the voiced regions. However, to maintain consistency in features, both of these sets of features were extracted only from the voiced region. The detection of epoch locations is described in Sect. 2.1. This is followed by the extraction of the epoch-based features corresponding to the epoch locations in Sect. 2.2. Finally, the MFCC features and the complementary analysis of the MFCC and epoch-based features are described in Sect. 2.3.

2.1 Extraction of the Epoch Locations

The following steps describe the detection of the epoch locations:

1. The voiced segment is detected using the phase of zero-frequency filtered signal [16].
2. The voiced speech signal is differentiated to emphasize the high-frequency components using the formula:

$$y[n] = s[n] - s[n - 1] \quad (1)$$

where $y[n]$ is the differentiated signal at the n th sample, $s[n]$ is the actual speech signal at the n th sample, and $s[n - 1]$ is the actual speech signal at the $(n - 1)$ th sample.

3. The differentiated speech signal is segmented into 3 msec frames and sampled at the rate of 16 kHz, resulting in $M = 48$ samples. These are appended with $N - M$ (2048-48) zeros to obtain sufficient frequency resolution, where N denotes the window length for short-time Fourier transform.
4. The time-domain signal is multiplied by the square of a window function h_1 (defined below) to smoothen the spectrum.

$$h_1[n] = \begin{cases} 0 & n = 0 \\ h_1[n] = \frac{1}{4\sin^2(\frac{\pi n}{N})} & n = 1, 2, \dots, N - 1 \end{cases} \quad (2)$$

where N denotes the window length as defined before.

The resulting windowed signal $x[n]$ is computed as:

$$x[n] = y[n] \times h_1[n]^2 \quad (3)$$

where $y[n]$ and $h_1[n]$ are as defined in Eqs. 1 and 2, respectively.

5. The spectral features are then emphasized by taking only the numerator of the group delay function of the windowed signal. The obtained output is called numerator group delay (NGD) spectrum and is computed as:

$$g[k] = X_R[k]Z_R[k] + X_I[k]Z_I[k], k = 0, 1, 2, \dots, N - 1 \quad (4)$$

where $X(k) = X_R[k] + X_I[k]$ is the discrete-time Fourier transform (DTFT) of $x[n]$ and $Z(k) = Z_R[k] + Z_I[k]$ is the DTFT of $z[n] = nx[n]$. The NGD spectrum is further differentiated two times to remove the trend. The resulting spectrum is called differentiated numerator group delay (DNGD) spectrum.

6. To prominently highlight the spectral peaks, Hilbert envelope of the DNGD spectrum is computed. For DNGD spectrum $g[k]$, its Hilbert envelope, denoted $h_e[k]$, is computed as:

$$h_e[k] = \sqrt{g^2[k] + g_h^2[k]} \quad (5)$$

where $g_h[k]$ is the Hilbert transformation of the DNGD spectrum $g[k]$.

The Hilbert transformation of the DNGD spectrum is called HNGD.

7. From each HNGD spectrum, the sum of three prominent peaks is taken. The resultant output is called a spectral energy profile. The glottal closure instants in the spectral energy profile show high SNR compared to the neighboring instants. Further, the spectral energy profile is smoothed using the five-point mean smoothing filter to remove the high-frequency components.
8. The spectral energy profile is convolved with a Gaussian filter; the size of this filter is determined using the average pitch period of the corresponding segment. A Gaussian filter of length L is given by

$$G[n] = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{n^2}{2\sigma^2}}, n = 1, 2, \dots, L \quad (6)$$

The standard deviation σ used in the above formula is taken as $\frac{1}{4}^{th}$ of the Gaussian filter length. The resulting output $g(n)$ is called epoch evidence [41].

9. The false peaks in the epoch evidence plot are removed using the following criteria:
 - (a) If two successive peaks having a difference of less than 2 ms are found, the peak with less amplitude is removed. This is because 2 ms is the minimum range of the pitch period.
 - (b) Successive actual peaks must bound a negative region between them. If a peak does not bound a negative region with the previous actual peak, it is considered spurious.
10. The positive peaks in the resulting epoch evidence plot represent epoch locations.

Figure 1 shows epoch detection using the ZTW method. The angry speech segment and its corresponding differentiated electroglottograph (dEGG) signal are shown in Fig. 1a, b, respectively. The spectral energy profile obtained from the HNGD spectrum of the speech signal using the ZTW method is plotted in Fig. 1c. The epoch evidence plot obtained after convolving the spectral energy profile with a Gaussian window is shown in Fig. 1d. The epoch locations are shown in Fig. 1e. The ZTW method for the detection of the epoch locations has been shown to be robust for emotional speech [41].

2.2 Epoch-Based Features

The epoch-based features such as instantaneous pitch, strength of excitation, and phase of the spectral energy profile are specific to each emotion [15]. These features have been used in the proposed model. Instantaneous pitch and strength of excitation are derived from epoch locations, while the phase features are derived from the spectral energy profile. The epoch locations and spectral energy profile are obtained by the ZTW method [41]. The advantage of this method is that the energy value at an epoch location is actually the sum of the glottal formants. Therefore, both time information and spectral information are retained at the epoch locations. The above-mentioned epoch-based features are described next.

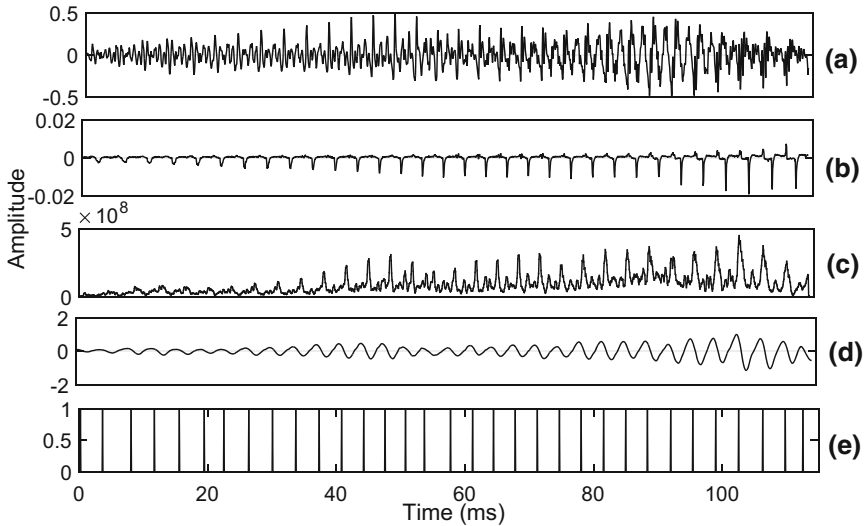


Fig. 1 Epoch detection using the ZTW method. **a** Angry speech segment. **b** dEKG signal. **c** Spectral energy profile obtained from the HNGD spectrum. **d** Epoch evidence plot. **e** Epoch locations

2.2.1 Instantaneous Frequency

Instantaneous period (IP) is the duration between two successive epoch locations. Instantaneous frequency, denoted Δf , is computed as the reciprocal of IP [14,25]:

$$\Delta f(i) = \frac{1}{t(i) - t(i+1)} \quad (7)$$

where $t(i)$ represents the i^{th} epoch location at time t .

2.2.2 Strength of Excitation (SOE)

This SOE is computed as the difference between the strengths of two successive epoch locations [9]:

$$\Delta y(i) = e(i) - e(i+1) \quad (8)$$

where $e(i)$ is the epoch strength at the i^{th} epoch location.

2.2.3 Instantaneous Phase

The phase of the spectral energy profile is obtained using the cosine of the phase function of the analytical signal. The spectral energy profile is used to derive the analytical signal. After that, the phase at each of the epoch locations is used as one of the epoch-based features.

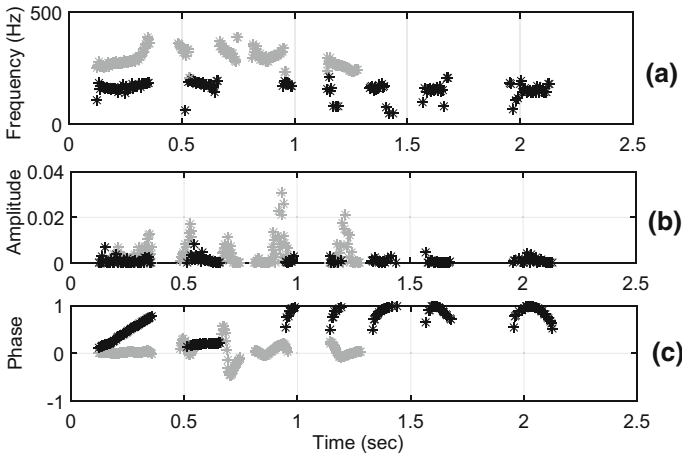


Fig. 2 Instantaneous pitch and SOE contours of angry (grey) and sad (black) speech signal using the proposed method. **a** Instantaneous pitch contour, **b** SOE contour, and **c** Instantaneous phase contour of angry and sad speech signals

- The analytic signal $g_a(n)$ corresponding to the spectral energy profile $g(n)$ is given by

$$g_a(n) = g(n) + jg_h(n) \tag{9}$$

where $g_h(n)$ is the Hilbert transformation of $g(n)$.

- The Hilbert envelope of the spectral energy profile $g(n)$, denoted $h_e[n]$, is calculated as:

$$h_e[n] = \sqrt{g^2[n] + g_h^2[n]} \tag{10}$$

where $g_h(n)$ is the Hilbert transformation of $g(n)$.

- The cosine of the phase of the analytic signal $g_a(n)$, denoted $\cos\Phi(n)$, is given by

$$\cos\Phi(n) = \frac{Re(g_a(n))}{|g_a(n)|} = \frac{g(n)}{h_e[n]} \tag{11}$$

where $g(n)$ is the spectral energy profile derived from the speech signal using the ZTW method and $Re(g_a(n))$ denotes the real part of $g_a(n)$.

The instantaneous frequency and SOE values of a speech utterance by a common speaker in different emotions are plotted in Fig. 2. Figure 2a shows the instantaneous pitch for two emotions: angry and sad. The gray color indicates angry emotion, while black indicates sad emotion. It is clear from Fig. 2a that the range of instantaneous pitch varies from 250 to 400 Hz for angry emotion and 100 to 200 Hz for sad emotion. Figure 2b shows the SOE values for two emotions: anger and sadness. The variation in

the SOE values is higher in the case of angry emotion than sad emotion. The variation in the SOE values is much less in the case of sad emotion. Figure 2b shows the phase of the spectral energy profile; it is higher for sad than for angry. The probability densities for each epoch-based feature are plotted in Fig. 3. The first four sessions of the IEMOCAP dataset are used to plot the probability densities. Figure 3a shows the probability density of instantaneous F0. The probability densities of neutral and sad emotions are almost the same, but the mean and variance of the probability densities of anger and happy emotions are different. Figure 3b shows the probability density of strength of excitation feature. Here also, the same observations as Fig. 3a can be made regarding the probability densities of different emotions. Figure 3c shows the probability densities of the instantaneous phase. The probability densities of happy and sad emotions can be easily discriminated from anger and neutral emotions. The probability densities of anger and neutral emotions are almost the same, but the mean and variance of probability densities of happy and sad emotions are different. The sequential information is lost in the probability densities plot; however, each emotion has its own temporal sequence. The sequential information of epoch-based features can be well captured with a dynamic model like the hidden Markov model (HMM).

2.3 Complementary Analysis of MFCC and Epoch-Based Features

The MFCC features also have emotion-specific information. The mel scale is used to mimic the behavior of the human auditory system, which gives high resolution at lower frequencies. These features are obtained by applying discrete cosine transform on the log power spectrum of the short-time speech signal. We combine the MFCC features with the epoch-based features in our model for recognizing emotions. The speech signal is processed frame-wise; the frame size is 20 ms with 10 ms overlapping. Thirteen MFCC features are obtained from each frame. The MFCC features are extracted using the method given in [30]. The recording variations are minimized by subtracting cepstral mean and normalizing the variance of the MFCCs at the utterance level.

Canonical correlation analysis (CCA) [13] is performed to show the complementary nature of the MFCC and epoch-based features. The CCA is performed between the 13 MFCC and three epoch-based features. Figure 4 shows the canonical correlation coefficients. There are three indices because the CCA gives the output dimension, which is the minimum of the dimensions of the MFCC and epoch-based features. The canonical correlation coefficients are very low, except for the first index. The value of the first index is high because the first index represents magnitude in both sets of features. The schematic diagram of the proposed feature combination and transformation is shown in Fig. 5.

3 Proposed Emotion Recognition System

An SER system is an outcome of two principal stages. In the first stage, training is performed using the features extracted from known emotional speech utterances. In

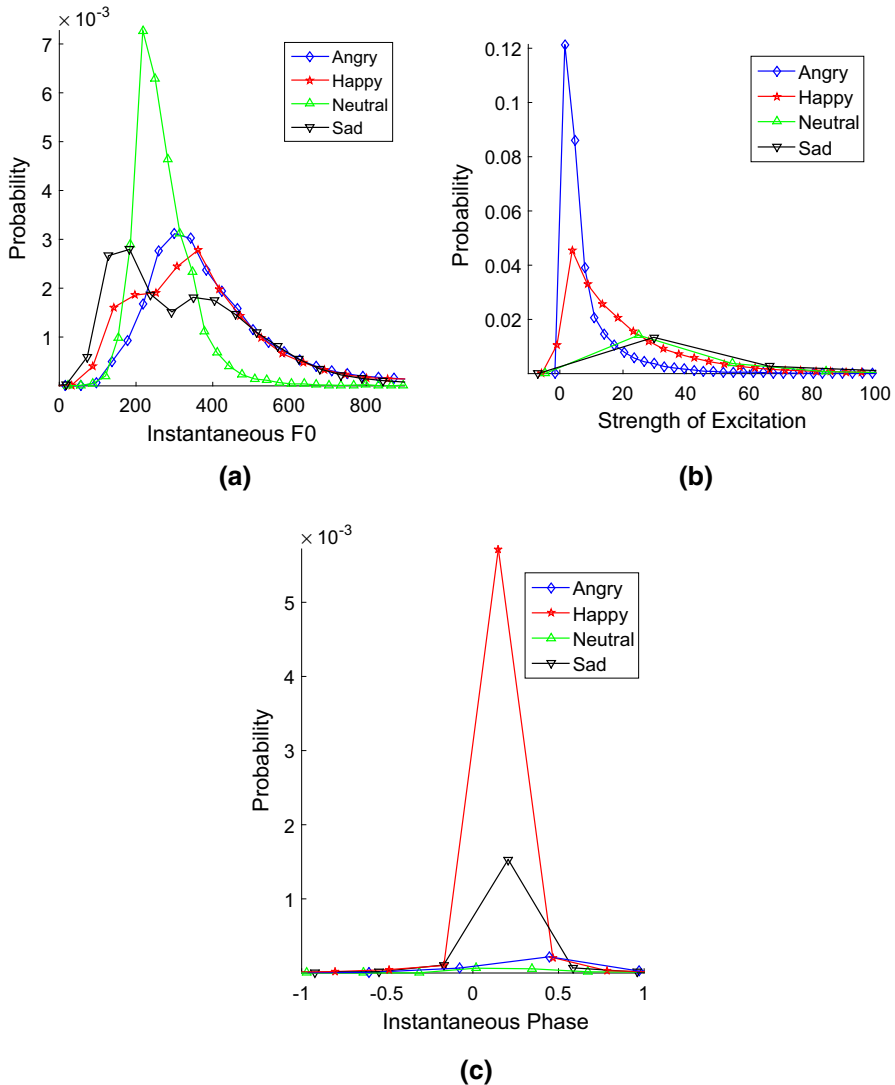


Fig. 3 Probability densities of instantaneous F0, strength of excitation, and instantaneous phase are plotted in a–c respectively

the second stage, i.e., the testing phase, the evaluation of the trained model is carried out on unseen emotional speech utterances. The schematic diagram of the proposed emotion recognition system is shown in Fig. 6.

We have combined the MFCC features with epoch-based features, namely instantaneous pitch, instantaneous phase, and strength of excitation. The epoch-based and MFCC features contain complementary information for recognizing emotions. Hence, the combined features significantly improve the accuracy of emotion recognition. After combining these feature vectors, linear discriminant analysis (LDA) and maximum

Fig. 4 Bar graph showing the canonical correlation coefficients between the MFCC and epoch-based features

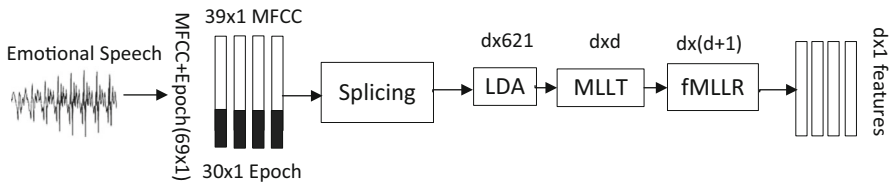
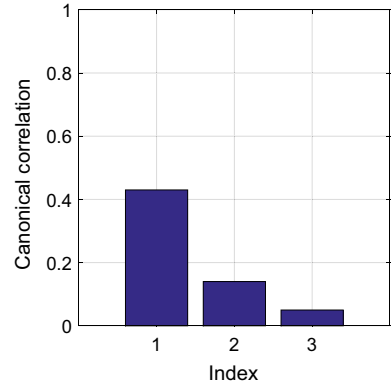


Fig. 5 Feature combination and transformation. Thirty-nine MFCC features and 30 epoch-based features are combined. For the resultant 69 features, nine frames are spliced to preserve the contextual information. The resulting 621 ($69 \text{ features} \times \text{nine frames}$) features are decorrelated using LDA and MLLT transform. d is the size of the transformed feature vector. Finally, fMLLR is used for speaker adaptation

likelihood linear transform (MLLT) are applied to decorrelate the feature vector; this enhances the accuracy of the model. fMLLR technique is then used for speaker-adaptive training (SAT); this further improves the accuracy of the model.

3.1 DNN-HMMs

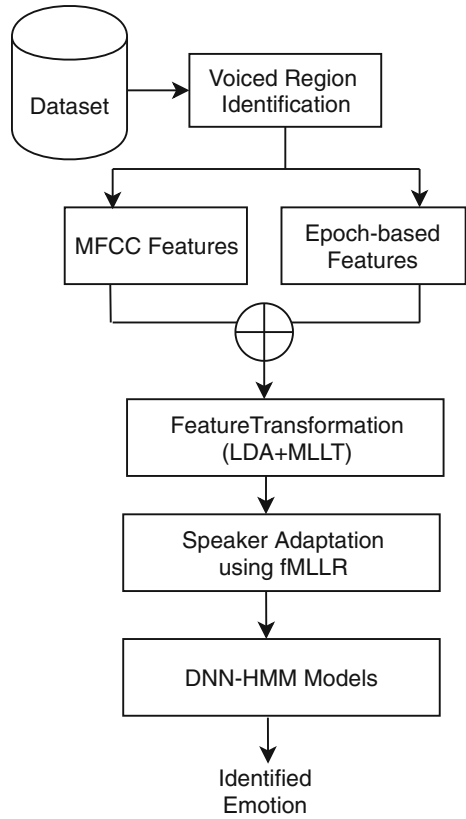
We have developed the SER system using HMM. It is a dynamic modeling approach that captures the temporal dynamic characteristics [28] of the transformed features of the corresponding emotions. In a conventional HMM, the observation probabilities of HMM states are estimated by GMMs (Gaussian mixture models). The GMMs used in such conventional HMMs are statistically inefficient to model nonlinear data in the feature space [18], whereas DNNs are capable of modeling nonlinear data. Therefore, GMMs are replaced with a DNN to estimate the observation probabilities of observing an input sequence at each state in the training phase. This approach is called DNN-HMM.

3.1.1 GMM-HMM

An HMM model is defined as a quartet $\lambda = (R, A, B, \pi)$ comprising the following components:

- (a) $R = r_1, r_2, \dots, r_Q$ denotes the set of Q hidden states.

Fig. 6 Schematic diagram of the proposed SER system



(b) A denotes the transition probabilities for states in R . $a_{jk} \in A$ is defined as

$$a_{jk} = P(q_{t+1} = r_k | q_t = r_j), 1 \leq j, k, \leq |Q| \tag{12}$$

where q_t denotes the state at time t .

(c) B denotes the set of observation probabilities. $b_j(i_t) \in B$ is defined as the probability of observing the input frame i_t in the state r_j . B is constructed by a finite number of mixture components L .

$$b_j(i_t) = \sum_{l=1}^L \phi_{jl} \mathfrak{N}(j_t, \mu_{jl}, C_{jl}), 1 \leq j \leq |Q| \tag{13}$$

where ϕ_{jl} is the weight of the mixture component for the l_{th} mixture in the state r_j and \mathfrak{N} is a Gaussian function with mean vector μ_{jl} and covariance matrix C_{jl} for the l_{th} mixture component in the state j_t at time t . Due to the use of a Gaussian function, such an HMM is called GMM-HMM.

(d) π denotes the initial state probabilities. $\pi_j \in \pi$ is defined as:

$$\pi_j = P(q_1 = r_j), 1 \leq j \leq |Q| \quad (14)$$

where r_j is the hidden state and Q is the number of hidden states.

The training and testing procedure in HMM is as follows:

1. **Training Procedure** Given a set of training data X , the model parameters are obtained using the forward–backward algorithm and Baum–Welch algorithm such that $\lambda^* = \operatorname{argmax}_\lambda P(X|\lambda)$.
2. **Testing Procedure** We have a model λ and a test sequence $I = (i_1, i_2, \dots, i_T)$. The testing problem is formulated as finding the optimal hidden state sequence (q_1, \dots, q_T) that has most likely attained the test sequence I . This is achieved by the Viterbi algorithm as follows

$$P(\lambda|I) = \max_{q_1, \dots, q_T} \pi_{q_1} \prod_{t=2}^T P(q_t|q_{t-1})b_{q_t}(i_t) \quad (15)$$

where π_{q_1} , $P(q_t|q_{t-1})$ and $b_{q_t}(i_t)$ are the initial state, transition state, and observation state probabilities, respectively.

We built N HMMs λ_n , ($n = 1, \dots, N$) for N different emotion classes. A new utterance I is assigned an emotional class using:

$$n^* = \operatorname{argmax}_{1 \leq n \leq N} P(\lambda_n|I) \quad (16)$$

where $P(\lambda|I_n)$ is obtained using the Viterbi algorithm as in Eq. 15.

3.1.2 DNN-HMM

The procedure followed for training and testing in DNN-HMM is as mentioned in [18]]. The steps in detail are as follows:

1. Training Procedure

- (a) A GMM-HMM λ_n with Q states is trained for each emotion class using the training occurrences of that class, where $n = 1, \dots, N$ (N denotes the number of emotional classes).
- (b) For every utterance in the training set $I = (i_1, i_2, \dots, i_T)$ for the n^{th} emotional class, Viterbi algorithm in Eq. 15 is applied on λ_n to find an optimal state sequence $(q_1^n, \dots, q_t^n, \dots, q_T^n)$. A label $L_i (i \in (1, \dots, N \times Q))$ is assigned to each state q_t^n by the state label mapping table.
- (c) The training utterances, combined with the corresponding labeled state sequences, are then fed as input to a DNN. The outputs of the DNN are the $N \times Q$ posterior probabilities corresponding to the output units.

2. Testing Procedure

- (a) The test feature sequence I is passed by the DNN to estimate the posterior probabilities $p(L_j|i_t)_{j=1,\dots,N \times Q}$ as outputs. After this, the posterior probability $p(q_t = S_k^n|i_t)$ is determined from $p(L_j|i_t)$ by aligning the label L_j to the state k of the class n with the help of the state-label mapping table.
- (b) The observation probability of each state, denoted $p(i_t|q_t)$, is obtained using Bayes' theorem as follows:

$$p(i_t|q_t) = \frac{p(q_t|i_t) * p(i_t)}{p(q_t)}, i_t \in I \quad (17)$$

The prior probability $p(q_t)$ is obtained from the occurrence of the training set, and the probability $p(i_t)$ remains constant because the input frames are assumed to be mutually independent.

- (c) For an unseen speech utterance, the probability of each emotion model λ_n is estimated using the Viterbi algorithm (as in Eq. 15). The utterance is assigned the class whose estimated likelihood probability $p(I|\lambda_n)$ is maximum; however, the observation probability is replaced by Eq. 17.

Four HMMs, corresponding to the four emotion classes, and a DNN were built. In the testing phase of DNN-HMM, the posterior probability of each state is calculated using DNN instead of GMM. For this, GMM-HMM is first applied to obtain the optimal sequence of states. Thereafter, a DNN model is developed that takes a combination of emotion and optimal states (four emotions \times five states) as output and the training dataset as input. This DNN network predicts the states of the given frame for testing data. The observation probability of each state is easily derived using the DNN prediction by Eq. 17.

3.2 Speaker Adaptation

Adaptation is a necessary task for emotion recognition. In general, we train our model with a limited dataset, but in a real environment, there may be different types of speakers and noise. One must have a robust method to adapt the trained model in a real environment. In our work, fMLLR transformation has been applied per speaker to adapt the emotion variation of different speakers.

3.2.1 Review of the fMLLR Approach

Feature-space transformation, also called constrained MLLR, fMLLR, or CMLLR, is a very popular technique in speech recognition for SAT [22].

The fMLLR transformation is performed as follows:

$$\hat{x}_i = Px_i + b = W\xi_i \quad (18)$$

where

x_i is the input feature vector to be transformed,

P is the rotation matrix and b is a bias term,

$W = [b \ P]$ is the $d \times (d + 1)$ transformation matrix (d is the size of the feature vector), and

$\xi_i = [1 \ x_i]$ is the extended feature vector.

In order to find the optimal value of W , the following likelihood function is maximized using the expectation-maximization (EM) approach:

$$Z(W; X, M) = \sum_i \log(p(W\xi_i; M)) \quad (19)$$

where

X represents the utterances in the dataset and

M is the GMM model for which fMLLR is performed.

3.2.2 Speaker-adaptive Training and Testing

1. **Training-time speaker adaptation** In general, the speaker information of the training dataset is available. This information can be effectively utilized during training. For example, out of the ten speakers in the dataset, eight are used for training and two for testing. Let $E(n)$ denote the set of n emotions under consideration.

A training process is as follows:

- A set of models $\{M(n)\}$ for each emotion in $E(n)$ is trained using randomly chosen speaker, say $s_1 \in \{s_i\}$, where $i \in \{1, \dots, 8\}$.
 - Next, for a randomly chosen speaker, say s_2 , from the remaining set of speakers $\{s_i\}$, where $i \in \{2, \dots, 8\}$, a set of transformations represented as $\{W_{s_2(n)}\}$ is estimated for each emotion in $\{E(n)\}$. Then, the corresponding models in set $\{M(n)\}$ are retrained using the respective transformed feature vector, re-estimating the parameters of the GMM models $M(n)$. This process is repeated for all the remaining speakers $\{s_i\}$, $i \in \{3, \dots, 8\}$ in the training set.
2. **Testing-time speaker adaptation** fMLLR is also used to transform the testing set of speakers $\{s_i\}$, $i \in \{9, 10\}$ so that their emotions can be identified by the emotion models of training-set speakers. For each emotion E_n in $\{E_n\}$, the emotion models $\{M(n)\}$ have already been built using the training-set speakers. For each testing-set speaker $\{s_i\}$, $i \in \{9, 10\}$, $\{W_{s_i(n)}\}$ is estimated using Eq. 19. The speaker-adaptive training and testing is performed on the LDA+MLLT transformed feature vectors.

4 Dataset

Our proposed model has been evaluated on the IEMOCAP dataset [3]. It is a multi-modal dataset that contains audio, video, text, and gesture information of conversations arranged in dyadic sessions. The dataset is recorded with ten actors (five male and five female) in five sessions. In each session, there are conversations of two actors, one

from each gender, on two subjects. The conversation of one session is approximately five minutes long. The contents of the dataset are recorded in both scripted and spontaneous scenarios. The total number of utterances in the dataset are 10,039, out of which 4,784 utterances are from the spontaneous sessions and 5,225 are from the scripted sessions. The average length of an utterance is 4.5 seconds, and the average word count per utterance is 11.4 words. The total recording time of the dataset is about 12 hours. The dataset is labeled as per the two popular schemes: discrete categorical label (i.e., labeled as happy, angry, neutral, and sad) and continuous dimensional label (i.e., activation, valence, and dominance). For our experiments, we used the speech utterances and the corresponding discrete emotion labels.

5 Experimental Setup and Discussion of Results

We have developed a speaker-adaptive DNN-HMM-based SER system that combines both the MFCC and epoch-based features. The proposed framework has been evaluated on the IEMOCAP [3] dataset for four emotions: *angry*, *happy*, *neutral*, and *sad*, and compared with state-of-the-art techniques [12,20,23].

The contributions of our work are as follows:

1. **Use of the epoch-based features extracted by the ZTW method** We used the epoch-based features where the epoch locations are extracted using the zero-time windowing (ZTW) method. This method is robust for emotional speech [41]. Epoch-based features, namely instantaneous pitch, phase, and strength of excitation (SOE), were used. Further, the epoch-based features were combined with the MFCC features. The proposed approach showed (i) an improvement of +7.13% over state-of-the-art techniques and (ii) an improvement of +5.07% over MFCC features when speaker adaptation was used.

Novelty While epoch-based features have been used in the past for SER systems, this is the first time that the epoch-based features have been extracted using the ZTW method, which has been shown to be superior to its contemporary methods, especially for emotional speech [41].

2. **DNN-HMM Model** To capture the rapid variations in an emotional speech, we have developed a DNN-HMM model because HMMs are known to capture sequential changes in emotions [32]. The combined set of features (i.e., MFCC+Epoch-based) produced an improvement of +5.07% over the MFCC features when used in the proposed model.

3. **Use of speaker-adaptive DNN-HMM** A speaker-adaptive DNN-HMM model has been proposed, where the speaker adaptation is applied during both the training and testing phases through the use of fMLLR. The proposed speaker-adaptive model, when used with only the MFCC features, achieves an improvement of +2.06% over state-of-the-art techniques. This further increases to +7.13% when both the MFCC and epoch-based features are used along with the speaker adaptation.

Novelty Our proposed approach uses the fMLLR-based speaker adaptation technique, which, to the best of our knowledge, is the first use of this technique for SER systems.

- Improved performance** The proposed model achieved significant improvement over state-of-the-art techniques [12,20,23]. Using the MFCC and epoch-based features along with the speaker adaptive training, the proposed model achieved an average accuracy of 65.93%, an improvement of +7.13% over state-of-the-art techniques.

Three models were developed for emotion recognition: using system (MFCC) features, using source (epoch-based) features, and by combining the MFCC and epoch-based features. Speaker adaptation was employed to accommodate the variation in the acoustic features of different speakers. For speaker adaptation, fMLLR was used during both the training and testing phases. The four categorical (class) labeled emotions, namely angry, happy, sad, and neutral, have 1103, 595, 1084, and 1708 utterances, respectively, the total being 4490.

5.1 GMM-HMM Versus DNN-HMM

We have considered the following four emotions for experiments: angry, happy, sad, and neutral. We used the first four sessions (consisting of eight speakers) for training and the last session (consisting of two speakers) for testing. The sizes of the training and testing datasets are 3583 and 952 utterances, respectively. We empirically set the hyperparameter of DNN (i.e., the number of epochs, the number of layers, and the number of hidden nodes in a layer). The number of utterances in the test-set speakers ninth and tenth was 436 and 516, respectively. The test dataset is imbalanced because the number of samples in each emotion class is different. Hence, we calculated both weighted accuracy (WA) and unweighted accuracy (UWA). However, comparison with state-of-the-art techniques was made using UWA. It gives equal weight to all the classes, specifically to the minority classes.

We used MATLAB tool for feature extraction and KALDI toolkit [29] for developing the system. For the emotion recognition system developed using the MFCC features, 13 MFCC features were extracted from each frame. We also took the derivative and double derivative of the normalized MFCCs as features. Thus, the total number of MFCC features extracted for each frame was 39. Cepstral mean–variance normalization [36] was used at the utterance level to mitigate the recording variations.

To preserve the contextual information, we used the popular triphone model approach where each frame is spliced with the left four frames and the right four frames. Feature transformation was applied to the features from the nine spliced frames. These features were projected into a lower dimensional space using LDA. Further, MLLT [10,11] was used to decorrelate the resulting features to improve the results.

The emotion recognition system was similarly developed for epoch-based features. The three epoch-based features being used, namely instantaneous pitch, phase, and strength of excitation, were extracted using the ZTW method. We took frame of size 20 ms, same as the MFCC features, to extract the epoch-based features. The number of epoch-based features was different for each frame. To address the variation in the lengths of the epoch-based feature vectors, we fixed the length as 10: the maximum number of epochs encountered in any frame. If the size of a feature vector was less than 10, we padded the remaining length with zeros. Therefore, the total number of epoch-

Table 1 Average SER accuracy (%) using the MFCC features on the IEMOCAP dataset

Features	Model	UWA (%)
MFCC (monophone)	GMM-HMM	44.70
MFCC (triphone)	GMM-HMM	47.70
MFCC (LDA+MLLT)	GMM-HMM	51.25
MFCC (LDA+MLLT)	DNN-HMM	54.35
Epoch-based (LDA+MLLT)	DNN-HMM	50.55

The bold values denote the accuracy to highlight the best SER

based features per frame was 30 (ten epochs \times 3 features per epoch). We developed the DNN-HMM model for each emotion using these 30 epoch-based features. Finally, we combined the epoch-based and MFCC features, increasing the length of the feature vector to 69. Four HMMs, corresponding to the four emotion classes, and a DNN were built. The DNN architecture used was 40:512 \times 5:20, where 40 is the number of transformed input features to the DNN and 512 \times 5 represents 512 nodes in each of the five hidden layers. This DNN configuration was found to be optimal after experimenting with different-sized configurations. There were 20 output classes in the DNN model (20=4 \times 5, where 4 is the number of emotion classes and 5 is the number of hidden states in HMM). These output classes were treated as ground-truth and were obtained by GMM-HMM-based Viterbi algorithm. There were 18 Gaussian components in the GMM-HMM model. The initial learning rate of DNN was set to 0.005, and after 25 epochs, it was reduced to 0.0005. Additional 20 epochs were performed after this. The batch size used for training was 512.

We developed the baseline GMM-HMM system using (1) monophone training, (2) triphone training, and (3) triphone training with LDA+MLLT. We developed the DNN-HMM model with LDA+MLLT transformed features. Table 1 shows the SER accuracy using MFCC and its derivatives. We also applied LDA+MLLT transformation on MFCC and its derivatives. The triphone model gives a better result than the monophone model because it captures the contextual information. We estimated the observation probability using DNN instead of GMM as described previously in Sect. 3.1.2. The triphone model produces an improvement of 3% in emotion recognition accuracy over the monophone model as shown in Table 1. Our system gives the best results in the case of DNN-HMM. The average accuracy increases by 3.5% when the observation probability of HMM models is calculated by DNN instead of GMM as given in Table 1.

We similarly developed the DNN-HMM model for epoch-based features. The epoch-based features were transformed using LDA+MLLT and fed to the DNN-HMM model. The average SER accuracy for the model developed using the MFCC features is 54.35%. The average SER accuracy for the model developed using the epoch-based features is 50.55%.

5.1.1 Effect of Speaker Adaptation

We employed speaker-adaptive training and testing of the DNN-HMM model. In general, adaptation is applied during the testing phase. In this work, adaptation was

Table 2 Average SER accuracy (%) using the epoch-based, MFCC, and combined (MFCC+epoch-based) features on the IEMOCAP dataset

Features	UWA (%)
MFCC (LDA+MLLT)	54.35
Epoch-based (LDA+MLLT)	50.55
MFCC+Epoch-based (LDA+MLLT)	60.14
MFCC (LDA+MLLT+SAT)	60.86
Epoch-based (LDA+MLLT+SAT)	54.52
MFCC+Epoch-based (LDA+MLLT+SAT)	64.20

The bold values denote the SER accuracy using speaker adaptive training (SAT)

Table 3 Emotion recognition performance on the IEMOCAP dataset based on the MFCC features

MFCC features (average: 60.86)				
Emotion	Angry	Happy	Neutral	Sad
Angry	75.88	15.29	6.20	2.63
Happy	50.18	24.47	23.0	2.35
Neutral	8.13	3.96	67.18	20.73
Sad	0.64	1.45	22.0	75.91

The bold values denote the accuracy of the correctly classified emotions

Table 4 Emotion recognition performance on the IEMOCAP dataset based on the epoch-based features

Epoch-based features (average: 54.52)				
Emotion	Angry	Happy	Neutral	Sad
Angry	67.21	10.29	17.45	5.05
Happy	18.56	32.24	31.70	17.50
Neutral	15.23	14.40	53.71	16.66
Sad	3.00	6.05	24.00	64.95

The bold values denote the accuracy of the correctly classified emotions

applied during both the training and testing phases. There are two advantages of speaker adaptation during the training phase: (1) it reduces the variance among training speakers and (2) it ensures that there are enough samples present for each training speaker. fMLLR is used for speaker adaptation. The detailed description of speaker adaptation using fMLLR is discussed in Sect. 3.2.

There is a significant improvement in the recognition rate after applying speaker adaptation for the MFCC, epoch-based, and MFCC+epoch-based features. As given in Table 2, after applying fMLLR, the emotion recognition rate increases by 6.51%, 3.97%, and 5.79 % for the MFCC, epoch-based, and MFCC+epoch-based features, respectively.

The confusion matrix for the experiments carried out using the MFCC features with LDA+MLLT+SAT transformation on the DNN-HMM system is shown in Table 3. It can be observed that there is more confusion between the angry and happy emotions because both are high arousal emotions. The sad and neutral emotions also show confusion because both are low arousal emotions.

The confusion matrix in Table 4 shows the recognition performance for each emotion using the epoch-based features. From the experimental results, it can be concluded that the epoch-based features discriminate better between angry and happy emotions when compared to the MFCC features. This is due to the sequential nature of the epoch-based features and corroborates the observation made in Sect. 2 regarding the same. The accuracy of happy emotion is less compared to other emotions in either of the models. We believe that this is due to the less number of happy samples compared to other emotions in the IEMOCAP dataset.

5.2 Choosing the Fusion Scheme

In this work, the DNN network is trained using three fusion approaches, namely early, intermediate, and late fusion scheme, as proposed in [42]. In the early fusion scheme, the transformed MFCC and epoch-based features are combined and fed to the DNN. In the intermediate fusion scheme, two separate layers are defined, one each for the transformed MFCC and epoch-based features. In addition, there is a combination layer that combines the transformed MFCC and epoch-based features after learning their inherent characteristics. In the late fusion scheme, two separate DNN classifiers are used for the transformed MFCC and epoch-based features. The outputs of both classifiers are posterior probabilities, which are combined as the average of the posterior probabilities of each classifier. Five hidden layers are used for the DNN network. For the intermediate integration scheme, the best result is obtained with one separate layer and four combination layers.

The bar graph in Fig. 7 shows that emotion recognition accuracy is higher for the combined (MFCC+epoch-based) set of features than each feature set alone. It can also be seen that the intermediate fusion scheme is more reliable than the early or late fusion scheme. Table 2 describes the SER accuracy using the models developed using the MFCC, epoch-based, and MFCC+epoch-based features, respectively. The combined feature set improves the accuracy of emotion recognition by 5.07% for (LDA + MLLT + SAT) and 5.79% for (LDA + MLLT) compared to the corresponding models developed using the MFCC features. The average performance improvement is 5.07% over the MFCC features, which is the more accurate feature set among the two individual sets of features. This result proves that the system and excitation source features have complementary information for emotion recognition. Finally, the MFCC and epoch-based features are combined and transformed using LDA+MLLT transformation. The transformed features are fed as input to the DNN-HMM model.

We would like to point out that while we have explored early, intermediate, and late fusion approaches, Krothapalli and Koolagudi [15], who have also combined epoch-based and MFCC features, have only applied a late fusion approach at the model level.

5.3 Proposed Framework Versus State-of-the-art approaches

We have compared our proposed framework with the existing state-of-the-art approaches [12,20,23] evaluated on the IEMOCAP dataset. In [20], the authors showed

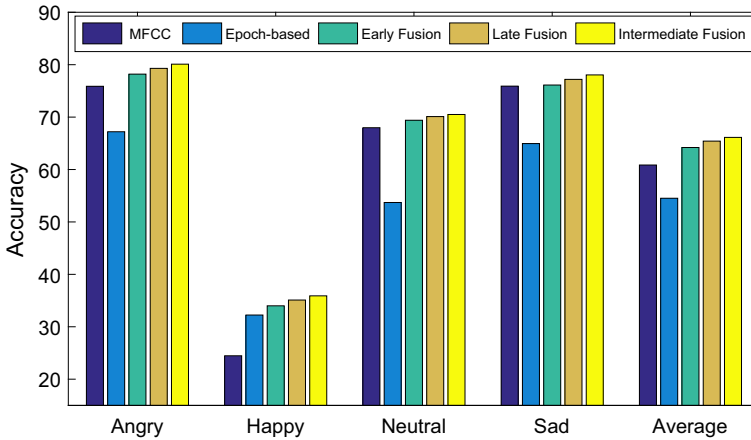


Fig. 7 Emotion classification performance (%) using the epoch-based, MFCC, and combined (MFCC+epoch-based) features on the IEMOCAP dataset

Table 5 Comparison of the proposed approach with state-of-the-art results on the IEMOCAP dataset

Model	Features	WA (%)	UWA (%)
SVM [20]	MFCC, Spectral features, Voice probability etc.	56.75	–
DNN+ELM [12]	MFCC, Pitch, and their derivatives	54.3	48.00
LSTM+attention [23]	Local-level descriptor, spectrogram	63.5	58.8
DNN-HMM (Proposed)	MFCC	64.3	60.86
DNN-HMM (Proposed)	Epoch-based	58.60	54.52
DNN-HMM (Proposed)	MFCC+Epoch-based	70.5	65.93

The bold values denote the accuracy of the proposed framework

the dependency of various features with speakers, text, and emotions. Most of the features that are used for emotion recognition vary with the variation of speakers and text. They normalized the speaker and text factors, after which the SER accuracy improved. In our work, speaker adaptation is applied instead of speaker normalization. A DNN-HMM model is developed for emotion recognition. This model captures the temporal sequence of emotion. This information is useful for differentiating among the same arousal emotions. In [12], a DNN-ELM model is developed in which a DNN model is used to extract the segment-level features. After that, the statistics of segment-level features are used as utterance-level features for the ELM classifier. This method used neither any technique to minimize the variance among speakers nor any model to capture the temporal sequence of emotion.

In [23], an LSTM model is used to capture the temporal sequence of the emotion. This work also used the attention method with the assumption that not all frames contain the same amount of contribution for classification. However, this method also does not use any techniques to minimize the variance among speakers.

In our framework, we used speaker adaptation to minimize the variance among speakers. As can be seen in Table 5, both the WA and UWA values of our proposed

framework outperform the other methods. The UWA accuracy of the proposed emotion recognition system using only MFCC features is 2.06 % more than the state-of-the-art results. The UWA accuracy improved by 7.13% with respect to state-of-the-art techniques [12,20,23] when the MFCC and epoch-based features are combined, as shown in Table 5.

6 Summary and Conclusion

This paper highlights the importance of speaker adaptation and complementary nature of the MFCC and epoch-based features. A DNN-HMM model is developed for each emotion using the MFCC features, epoch-based features, and a combination of MFCC+epoch-based features. The average emotion recognition rate of the proposed model using only the MFCC features is 60.86%—an improvement of 2.06% over the state-of-the-art techniques. The model developed using the MFCC features is further combined with the model developed using the epoch-based feature vectors. The observed accuracy of the combined model is 65.93%—an improvement of 7.13% over the state-of-the-art approaches. Based on these results, it may be concluded that the epoch-based features contain information complementary to the MFCC features for the emotion classification task. Our future work is to use the LSTM network to capture the contextual information of epoch-based features for emotion recognition.

Acknowledgements Akshay Deepak has been awarded Young Faculty Research Fellowship (YFRF) of Visvesvaraya PhD Programme of Ministry of Electronics & Information Technology, MeitY, Government of India. In this regard, he would like to acknowledge that this publication is an outcome of the R&D work undertaken in the project under the Visvesvaraya PhD Scheme of Ministry of Electronics & Information Technology, Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia).

References


1. D.O. Bos, EEG-based emotion recognition. *Infl. Vis. Audit. Stimul.* **56**(3), 1–17 (2006)
2. F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, B. Weiss, A database of German emotional speech, in *9th European Conference on Speech Communication and Technology* (2005)
3. C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **42**(4), 335 (2008)
4. C. Busso, A. Metallinou, S.S. Narayanan, Iterative feature normalization for emotional speech detection, in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5692–5695 (2011)
5. R.A. Calix, G.M. Knapp, Actor level emotion magnitude prediction in text and speech. *Multimed. Tools. Appl.* **62**(2), 319–332 (2013)
6. C. Clavel, I. Vasilescu, L. Devillers, G. Richard, T. Ehrette, Fear-type emotion recognition for future audio-based surveillance systems. *Speech Commun.* **50**(6), 487–503 (2008)
7. F. Dellaert, T. Polzin, A. Waibel, Recognizing emotion in speech, in *Proceeding of Fourth International Conference on Spoken Language Processing ICSLP'96*, vol. 3. IEEE, pp. 1970–1973. (1996)
8. F. Eyben, A. Batliner, B. Schuller, Towards a standard set of acoustic features for the processing of emotion in speech, in *Proceedings of Meetings on Acoustics 159ASA*, vol. 9. Acoustical Society of America, p. 060006 (2010)

9. P. Gangamohan, S.R. Kadiri, S.V. Gangashetty, B. Yegnanarayana, Excitation source features for discrimination of anger and happy emotions, in *15th Annual Conference of the International Speech Communication Association* (2014)
10. M.J. Gales, Maximum likelihood linear transformations for HMM-based speech recognition. *Comput. Speech Lang.* **12**(2), 75–98 (1998)
11. M.J. Gales, Semi-tied covariance matrices for hidden Markov models. *IEEE Trans. Speech Audio Process.* **7**(3), 272–281 (1999)
12. K. Han, D. Yu, I. Tashev, Speech emotion recognition using deep neural network and extreme learning machine, in *15th Annual Conference of the International Speech Communication Association* (2014)
13. D.R. Hardoon, S. Szedmak, J. Shawe-Taylor, Canonical correlation analysis: an overview with application to learning methods. *Neural Comput.* **16**(12), 2639–2664 (2004)
14. S.G. Koolagudi, R. Reddy, K.S. Rao, Emotion recognition from speech signal using epoch parameters, in *2010 international conference on signal processing and communications (SPCOM)*, pp. 1–5 (2010)
15. S.R. Krothapalli, S.G. Koolagudi, Characterization and recognition of emotions from speech using excitation source information. *Int. J. Speech Technol.* **16**(2), 181–201 (2013)
16. S.S. Kumar, K.S. Rao, Voice/non-voice detection using phase of zero frequency filtered speech signal. *Speech Commun.* **81**, 90–103 (2016)
17. C.M. Lee, S.S. Narayanan, Toward detecting emotions in spoken dialogs. *IEEE Trans. Speech Audio Process.* **13**(2), 293–303 (2005)
18. L. Li, Y. Zhao, D. Jiang, Y. Zhang, F. Wang, I. Gonzalez, E. Valentin, H. Sahli, Hybrid deep neural network–hidden Markov model (DNN-HMM) based speech emotion recognition, in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 312–317 (2013)
19. M. Mansoorizadeh, N.M. Charkari, Multimodal information fusion application to human emotion recognition from face and speech. *Multimed. Tools Appl.* **49**(2), 277–297 (2010)
20. S. Mariooryad, C. Busso, Compensating for speaker or lexical variabilities in speech for emotion recognition. *Speech Commun.* **57**, 1–12 (2014)
21. L. Mary, Significance of prosody for speaker, language, emotion, and speech recognition, in *Extraction of Prosody for Automatic Speaker, Language, Emotion and Speech Recognition*. Springer, Cham, pp. 1–22 (2019)
22. S. Matsoukas, R. Schwartz, H. Jin, L. Nguyen, Practical implementations of speaker-adaptive training, in *DARPA Speech Recognition Workshop* (1997)
23. S. Mirsamadi, E. Barsoum, C. Zhang, Automatic speech emotion recognition using recurrent neural networks with local attention, in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2227–2231 (2017)
24. R. Nakatsu, J. Nicholson, N. Tosa, Emotion recognition and its application to computer agents with spontaneous interactive capabilities. *Knowl.-Based Syst.* **13**(7), 497–504 (2000)
25. N.P. Narendra, K.S. Rao, Robust voicing detection and F_0 estimation for HMM-based speech synthesis. *Circuits Syst. Signal Process.* **34**(8), 2597–2619 (2015)
26. J. Nicholson, K. Takahashi, R. Nakatsu, Emotion recognition in speech using neural networks. *Neural Comput. Appl.* **9**(4), 290–296 (2000)
27. K.E.B. Ooi, L.S.A. Low, M. Lech, N. Allen, Early prediction of major depression in adolescents using glottal wave characteristics and teager energy parameters, in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4613–4616 (2012)
28. D. O’Shaughnessy, Recognition and processing of speech signals using neural networks. *Circuits Syst. Signal Process.* **38**(8), 3454–3481 (2019)
29. D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, The Kaldi speech recognition toolkit, in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (No. CONF)*. IEEE Signal Processing Society (2011)
30. L. Rabiner, Fundamentals of speech recognition. *Fundam. Speech Recognit.* (1993)
31. T.V. Sagar, *Characterisation and synthesis of emotions in speech using prosodic features*. Master’s thesis, Dept. of Electronics and communications Engineering, Indian Institute of Technology Guwahati (2007)
32. B. Schuller, G. Rigoll, M. Lang, Hidden Markov model-based speech emotion recognition, in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings (ICASSP’03)*, vol. 2. IEEE, pp. II–1 (2003)

33. B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, G. Rigoll, Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Trans. affect. Comput.* **1**(2), 119–131 (2010)
34. D. Ververidis, C. Kotropoulos, A state of the art review on emotional speech databases, in *Proceedings of 1st Richmedia Conference*, pp. 109–119 (2003)
35. D. Ververidis, C. Kotropoulos, I. Pitas, Automatic emotional speech classification, in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, pp. I-593 (2004)
36. O. Viikki, K. Laurila, Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Commun.* **25**(1–3), 133–147 (1998)
37. H.K. Vydana, S.R. Kadiri, A.K. Vuppala, Vowel-based non-uniform prosody modification for emotion conversion. *Circuits Syst. Signal Process.* **35**(5), 1643–1663 (2016)
38. Y. Wang, L. Guan, An investigation of speech-based human emotion recognition, in *IEEE 6th Workshop on Multimedia Signal Processing*, pp. 15–18 (2004)
39. C. Wu, C. Huang, H. Chen, Text-independent speech emotion recognition using frequency adaptive features. *Multimed. Tools Appl.* **77**(18), 24353–24363 (2018)
40. J. Yadav, K.S. Rao, Prosodic mapping using neural networks for emotion conversion in Hindi language. *Circuits Syst. Signal Process.* **35**(1), 139–162 (2016)
41. J. Yadav, M.S. Fahad, K.S. Rao, Epoch detection from emotional speech signal using zero time windowing. *Speech Commun.* **96**, 142–149 (2018)
42. D. Yu, L. Deng, *Automatic Speech Recognition*. Springer London Limited (2016)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Md. Shah Fahad¹  · Akshay Deepak¹ · Gayadhar Pradhan² · Jainath Yadav³

Akshay Deepak
akshayd@nitp.ac.in

Gayadhar Pradhan
gdp@nitp.ac.in

Jainath Yadav
jainath@cub.ac.in

¹ Department of Computer Science and Engineering, National Institute of Technology Patna, Patna, India

² Department of Electronics and Communication, National Institute of Technology Patna, Patna, India

³ Department of Computer Science, Central University of South Bihar, Gaya, India